Quarter 1 Project Report: Driver Liability Prediction Using Crash Data

Arjun Pagidi and Abhi Palikala

Statement/Project Goal

The goal of this project is to develop a machine learning model that predicts whether a driver was at fault in a traffic incident using the Montgomery County Crash Reporting Incidents dataset. By analyzing various factors such as driver behavior, environmental conditions, and crash characteristics, we aim to identify patterns and indicators that contribute to fault determination. This model can be utilized to improve road safety by highlighting key risk factors, guiding preventative measures, and informing traffic policies. Additionally, the insights from this project could be valuable for law enforcement agencies, insurance companies, and urban planners to better understand the dynamics of traffic incidents and ultimately reduce the number of preventable accidents on the road.

Description of Data Set

The data set consists of about 180,000 instances with 39 attributes (not including the class label).
Report Number: A unique identifier assigned to each crash report.
Local Case Number: Case number designated by the local law enforcement agency.
Agency Name: Name of the agency that reported the crash.
ACRS Report Type: Classification of the type of crash report (e.g., Standard, Supplement).
Crash Date/Time: The exact date and time when the crash occurred.
Route Type: Type of road where the crash took place (e.g., highway, local road).
Road Name: Name of the road where the incident occurred.
Cross-Street Name: Name of the nearest intersecting street.
Off-Road Description: Description of any off-road locations or areas impacted.
Municipality: City or town where the crash occurred.
Related Non-Motorist: Presence and type of non-motorist (pedestrian, cyclist) involved.
Collision Type: Type of collision (e.g., rear-end, side-impact).
Weather: Weather conditions at the time of the incident.
Surface Condition: Road surface condition (e.g., wet, dry, icy).
Light: Lighting conditions (e.g., daylight, dark, dawn).
Traffic Control: Presence and type of traffic control devices (e.g., stop signs, signals).
Driver Substance Abuse: Information on driver impairment due to substances.
Non-Motorist Substance Abuse: Information on non-motorist impairment due to substances.
Person ID: Unique ID for the person involved in the incident.
Driver At Fault: Indicates whether the driver was found to be at fault.
Injury Severity: Level of injury sustained (e.g., none, minor, fatal).
Circumstance: Additional circumstances contributing to the crash (e.g., speeding, swerving).
Driver Distracted By: Source of driver distraction (e.g., phone, passengers).
Drivers License State: State that issued the driver's license.
Vehicle ID: Unique ID for the vehicle involved in the crash.
Vehicle Damage Extent: The extent of vehicle damage (e.g., minor, total loss).
Vehicle First Impact Location: Initial point of impact on the vehicle (e.g., front, rear).
Vehicle Body Type: Type of vehicle body (e.g., sedan, SUV, truck).
Vehicle Movement: Movement of the vehicle at the time of the crash (e.g., turning, stopped).

Vehicle Going Dir: Direction the vehicle was moving (e.g., north, south).
Speed Limit: Speed limit of the road at the crash site.
Driverless Vehicle: Indicates if the vehicle was driverless.
Parked Vehicle: Indicates if the vehicle was parked at the time of the incident.
Vehicle Year: Year the vehicle was manufactured.
Vehicle Make: Manufacturer of the vehicle (e.g., Toyota, Ford).
Vehicle Model: Model of the vehicle (e.g., Camry, Mustang).
Latitude: Latitude coordinates of the crash location.
Longitude: Longitude coordinates of the crash location.
Location: General location description of the crash site.
Dataset:
https://catalog.data.gov/dataset/crash-reporting-drivers-data

Data Pre-Processing

The very first steps to our data preprocessing was to augment the CSV file in a way so that WEKA could open and understand the data, and so we may convert the CSV file into the ARFF format when desired. Much of the error resulted from several improperly formatted data points, whether that be a multi-line text block, a value using ‘’, or an ordered pair instead of a single value.

New Line

The new line issues arose from the ‘Off Road Description’ column originally, with several entries using the newline character, which WEKA was unable to make agreement with in a CSV file. Because all the values with newline characters were within this one column, and the column itself was so sparsely populated, (>90% of the data missing), we elected to remove the column entirely. This was again justified by the lack of relevance of the attribute (Off-Road Description), which was a qualitative, up-to the driver description of the location in which the crash occurred. This is not even tangentially related to whether the driver was at fault, resulting in its removal.

Quotation Marks and Single Quotes

The prevalence of ‘ and “ impeded WEKA’s ability to handle and understand the CSV file, so it was necessary that they were removed or replaced. To handle this, we utilized the ‘find and replace’ feature, which we used to remove these values from the dataset.

Date/Time

We believed that the time of day would be more useful than the day of the year when determining whether the driver was at fault, and we also believed that it would be easier to discretize. However, in order to discretize the data, we must first remove the date from the ‘Crash date/time’ attribute.

DateFormatting.py

Users > abhinav > Desktop > Programming > Python > DateFormatting.py > convert_to_minutes

```python
import csv
from datetime import datetime

def convert_to_minutes(date_time_str):
    dt = datetime.strptime(date_time_str, "%m/%d/%Y %I:%M:%S %p")

    minutes_since_midnight = dt.hour * 60 + dt.minute
    return minutes_since_midnight

input_file = 'Crash_Reporting_-_Drivers_DataV4.csv'
output_file = 'CrashReportingDateTime.csv'

column_name = 'Crash Date/Time'

with open(input_file, mode='r', newline='') as infile, open(output_file, mode='w', newline='') as outfile:
    reader = csv.DictReader(infile)
    fieldnames = reader.fieldnames
    writer = csv.DictWriter(outfile, fieldnames=fieldnames)

    writer.writeheader()

    for row in reader:
        row[column_name] = convert_to_minutes(row[column_name])
        writer.writerow(row)

```

The code follows the only column with the date and time, and converts it to minutes.
For an example of how this transformed our data, consult the image.

Crash Date/Time	Crash Date/Time
05/27/2021 07:40:00 PM	1180
09/11/2015 01:29:00 PM	809
	865
08/17/2018 02:25:00 PM	1080
08/11/2023 06:00:00 PM	1122
12/06/2023 06:42:00 PM	669
08/28/2023 11:09:00 AM	750
07/27/2023 12:30:00 PM	1000
12/29/2023 04:40:00 PM	1224
11/10/2023 08:24:00 PM	1173
10/16/2023 07:33:00 PM	634
09/30/2023 10:34:00 AM	

Removing Immediately Useless Attributes

The attributes that we concluded may be immediately removed from the dataset were ‘Report Number’ , ‘Local Case Number’, ‘Off Road Description’, ‘Municipality’, ‘Related Non-Motorist’, ‘Non-Motorist Substance Abuse’, ‘Person ID’, ‘Vehicle ID’, ‘Vehicle Model’, ‘Location’, ‘Circumstance’, and ‘Vehicle Make’. Several attributes such as report number, case number, person ID, and vehicle ID were removed due to being several long identifying strings

that were randomly generated. By definition, they should not have any predictive value. The attribute ‘vehicle model’, proved to be too inconsistent when being inputted, and seemed to be beyond the help of any data smoothing. Location was redundant itself with the presence of the attributes ‘latitude’ and ‘longitude’. Then, ‘Off Road Description’, ‘municipality’, ‘related non-motorist’, and ‘non-motorist substance abuse’ were very sparsely populated, with under 10% of the instances containing values. ‘Circumstance’ conveyed redundant information already given by ‘Surface Condition’, and we decided to remove ‘Vehicle Make’ because of the inconsistency of the values: there were several, different misspellings of ‘Toyota,’ for example. Finally, ‘Road Name’ had over 1000 different qualitative values, which would have made our classification algorithms too slow. This leaves us with 24 attributes excluding the class label to aid us with our prediction model.

Filling in missing values

The attributes with missing values include Route Type (18141, 10% missing), Road Name (18708, 10% missing), Surface Condition (17155, 9% missing), Traffic Control (1120, 1% missing), Injury Severity (789, 0% missing), Driver Distracted By (862, 0% missing), Drivers License State (11432, 6% missing), Vehicle First Impact Location (156, 0% missing), Vehicle Body Type (2597, 1% missing), Vehicle Movement (927, 1% missing), Vehicle Going Dr (4744, 3% missing), and Parked Vehicle (1534, 1% missing).
These were all replaced using WEKA’s built in unsupervised ‘ReplaceMissingValues’. There were no instances in which the class label was not filled in, so this was a valid step to take.

Discretize

The main target of discretization was the ‘Crash Date/Time’. By converting it to minutes, we hoped that discretizing it would provide a good sense of ‘early morning’, ‘midday’, ‘evening’ and ‘night’. Some numerical attributes, like "Latitude," were also binned. The discretization method varied, with some attributes using equal width and others using equal depth, depending on factors such as the presence of outliers and data distribution. This way, we were able to use classification algorithms even with initially continuous data.

Train Test Split

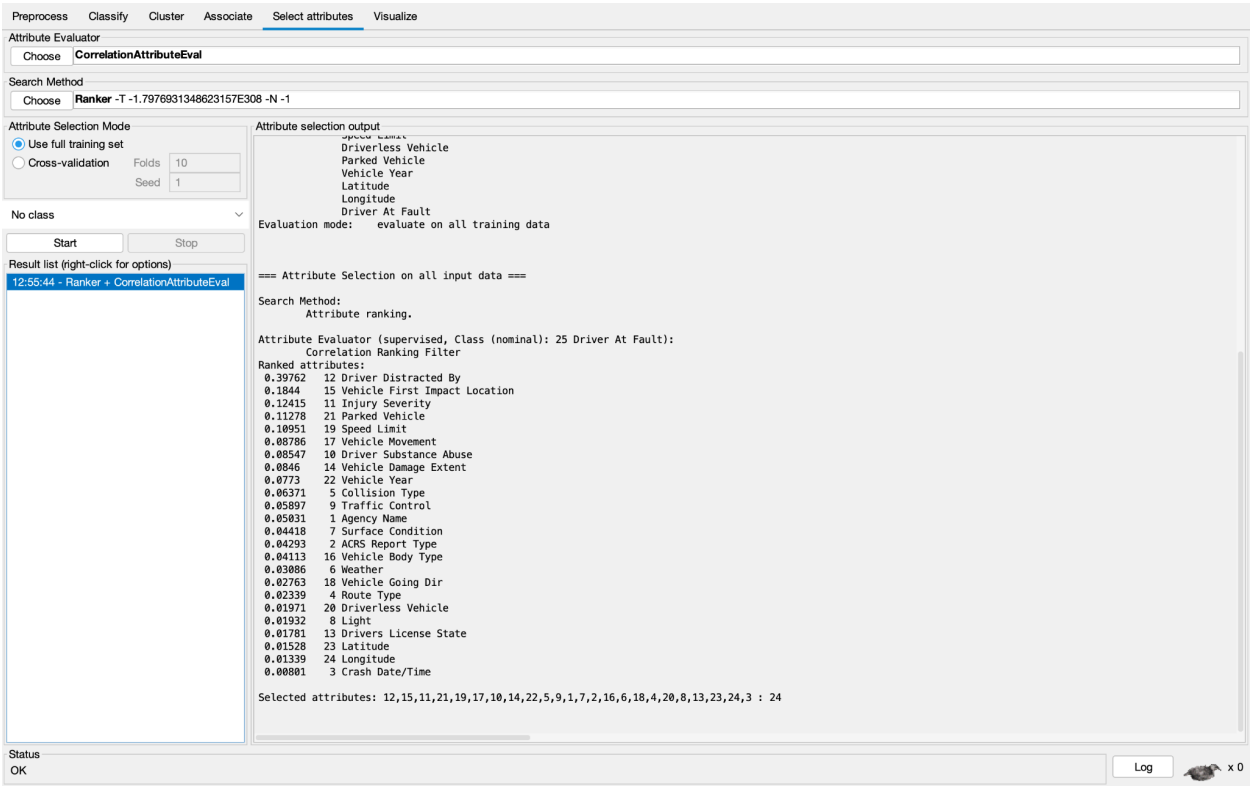
We did a stratified random train/test split with the use of Weka’s “stratifiedRemoveFolds”, yielding an 80-20 split stratified with respect to the class label “At Fault”. The resulting train and test splits had the same percentage of each class (to the nearest whole instance).

Attribute Selection

As stated previously, we had 25 attributes to work with to aid our prediction of 'Driver At Fault.' To narrow down the most useful attributes, we created five different datasets based on several different attribute selection algorithms and reasoning.

Correlation Based Feature Selection

Using the CorrelationAttributeEval attribute selection algorithm and the Ranker search method, we obtained the following analysis.

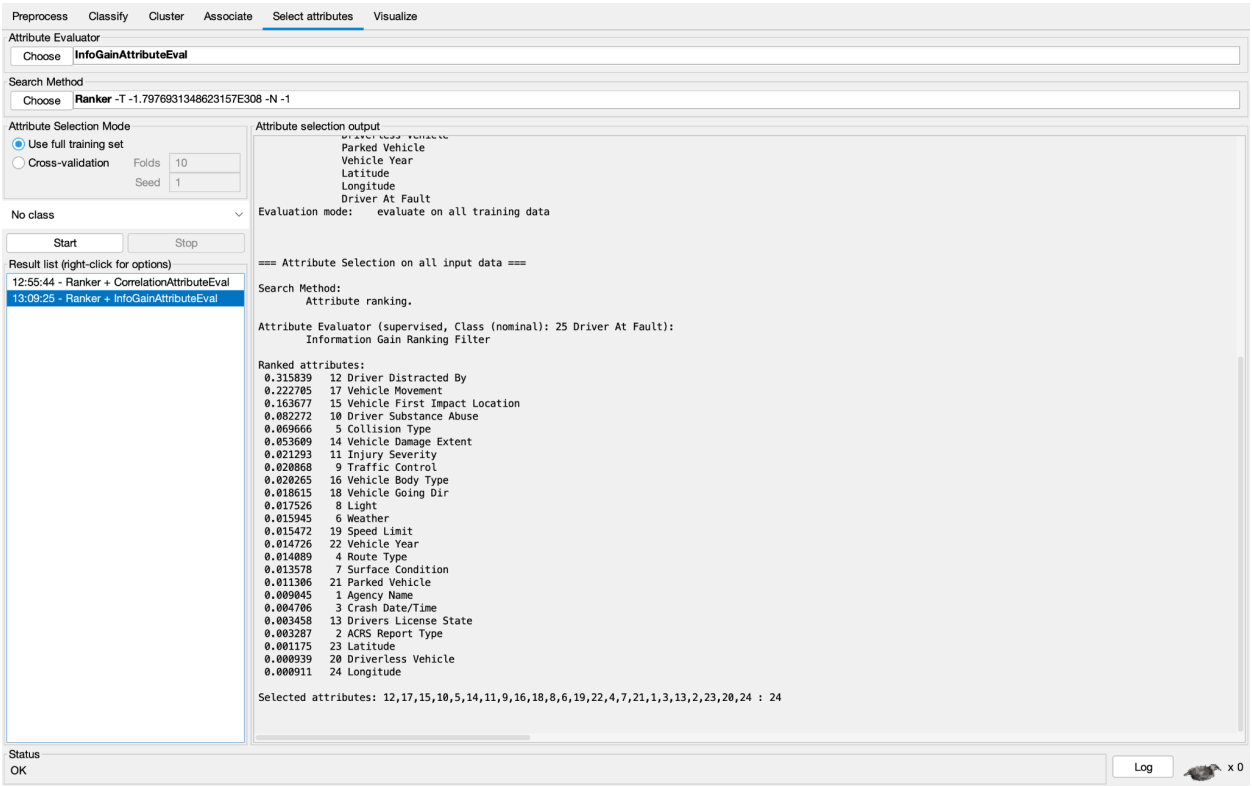


When selecting our attributes, we set a cutoff of 0.075. Thus, the following attributes were selected to be included based on this algorithm: 'Driver Distracted By', 'Vehicle First Impact Location', 'Injury Severity', 'Parked Vehicle', 'Speed Limit', 'Vehicle Movement', 'Driver Substance Abuse', 'Vehicle Damage Extent', and 'Vehicle Year'.

Info Gain Based Attribute Selection

Our second dataset was created using the InfoGainAttributeEval algorithm provided by WEKA, which also uses the Ranker search method.

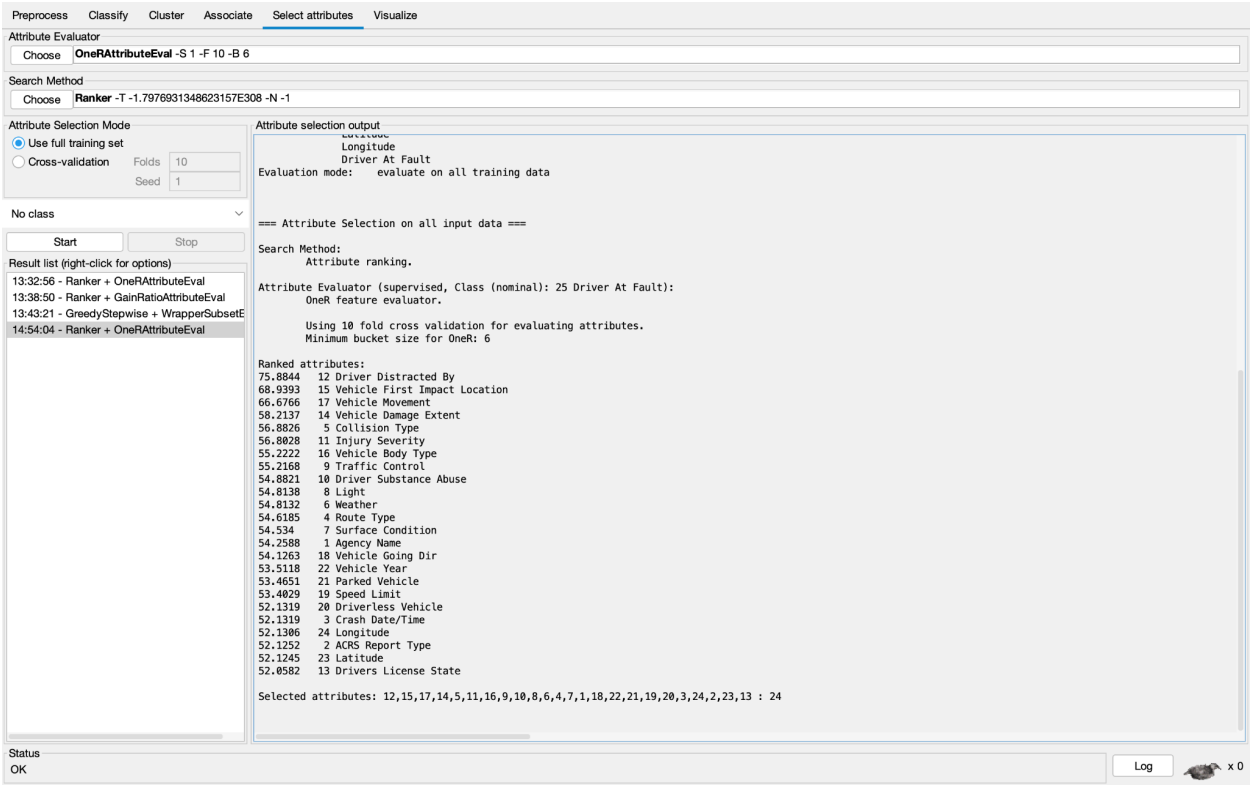
The results are shown below.



We chose a cutoff of 0.02. As such, we kept the attributes ‘Driver Distracted By’, ‘Vehicle Movement’, ‘Vehicle First Impact Location’, ‘Driver Substance Abuse’, ‘Collision Type’, ‘Vehicle Damage Extent’, ‘Injury Severity’, ‘Traffic Control’, and ‘Vehicle Body Type’.

One Attribute Based Selection

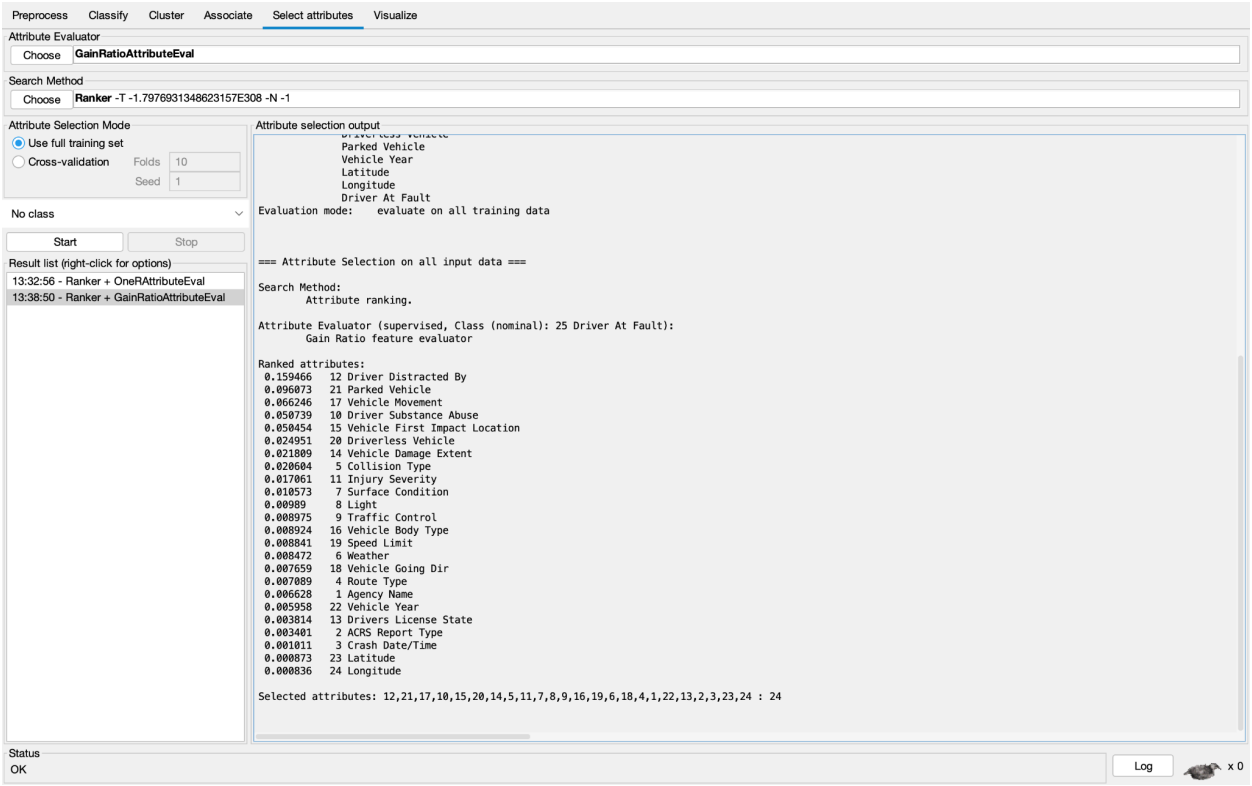
This dataset was created using the OneRAttributeEval attribute evaluator, which uses the Ranker search method. The following results were obtained.



Based on the cutoff value of 54.5, the attributes we chose to keep were: ‘Driver Distracted By’, ‘Vehicle First Impact Location’, ‘Vehicle Movement’, ‘Vehicle Damage Extent’, ‘Collision Type’, ‘Injury Severity’, ‘Vehicle Body Type’, ‘Traffic Control’, ‘Driver Substance Abuse’, ‘Light’, ‘Weather’, ‘Route Type’, and ‘Surface Condition’.

Gain Ratio Evaluation

This dataset was created using the GainRatioAttributeEval, which uses the Ranker search method. The results are shown below.



After setting a cutoff of 0.01, we decided to keep the following attributes: ‘Driver Distracted By’, ‘Parked Vehicle’, ‘Vehicle Movement’, ‘Driver Substance Abuse’, ‘Vehicle First Impact Location’, ‘Driverless Vehicle’, ‘Vehicle Damage Extent’, ‘Collision Type’, ‘Injury Severity’, and ‘Surface Condition’.

Self Selected Attributes
For the self-selected attributes, we chose to include any attribute that was suggested to be kept by *any* one of the attribute selection algorithms above. In other words, the only attributes removed were the ones that were considered 'useless' by every single attribute selection algorithm.

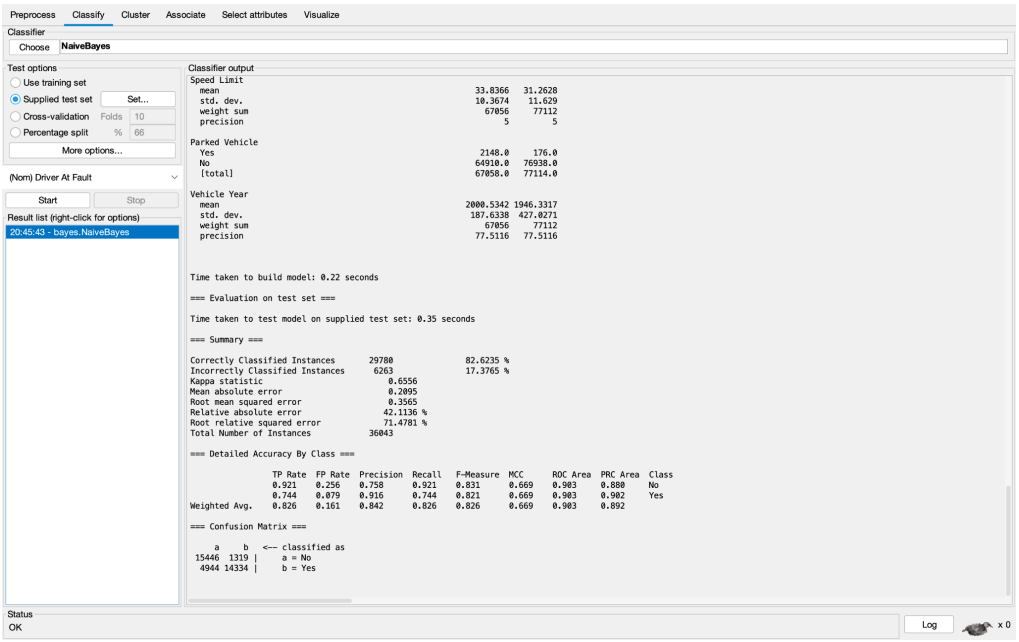
Performance Analysis of Models
The models we used were:

- J48
- Naive Bayes: This model forms probabilistic predictions based on the training set using Bayes' theorem and the naive assumption that attributes are independent from one another. The following formulas are used:

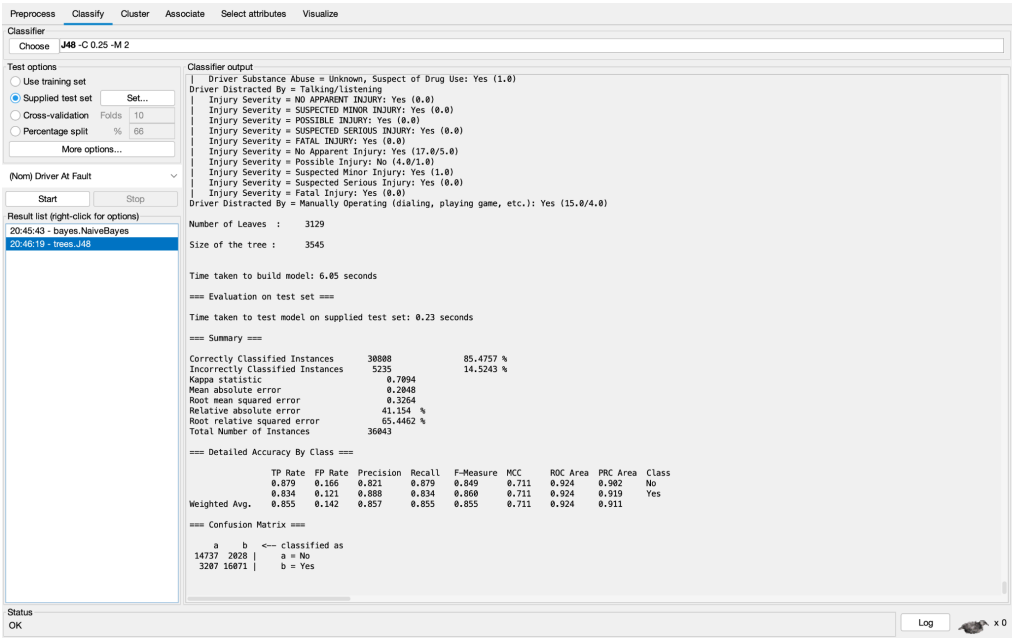
  $P(C_i|X) = \frac{P(X|C_i)\,P(C_i)}{P(X)}$ *when* $P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$ *where* $X = (x_1, x_2, \ldots, x_n)$ *is the instance. The prediction is the class* (C_i) *with the highest* $P(C_i|X)$ *value.*
- 1R: Level 1 decision tree. It forms a rule set with the following pseudocode:

  For each attribute
  For each value of the attribute
  count frequency of each class
  find the most frequent class
  make rule: assign that class to this attribute-value
  Compute the error rate of the rules (of this attribute)
  Choose the rules with the smallest error rate
- Random Tree

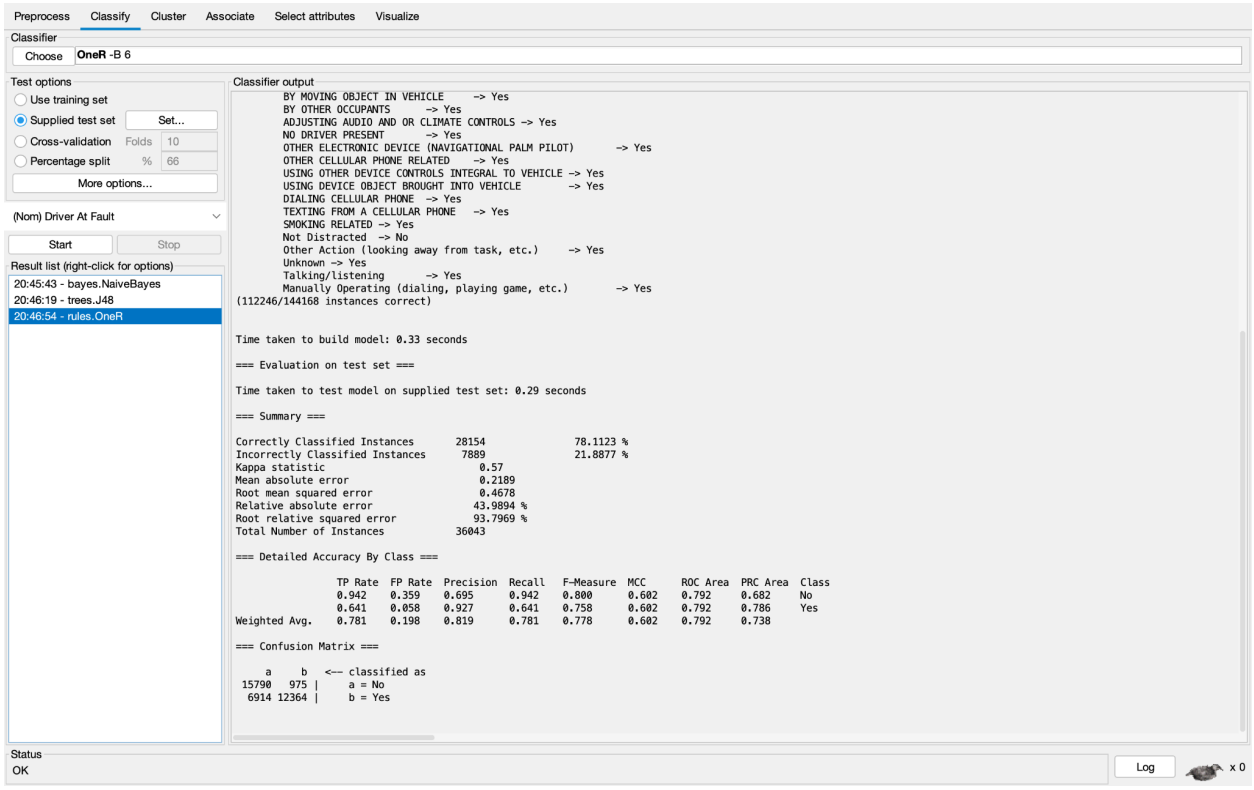
Results
Naive Bayes: CorrelationAttributeEval



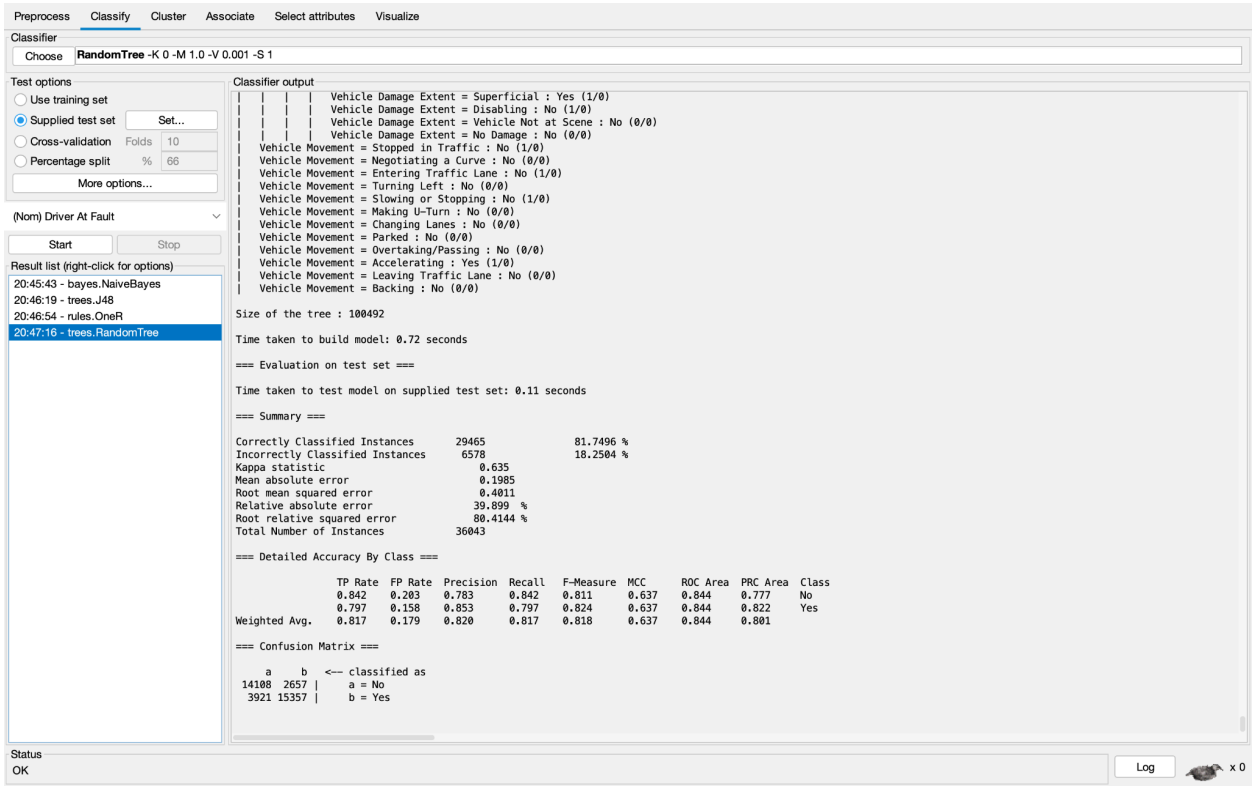
J48: CorrelationAttributeEval



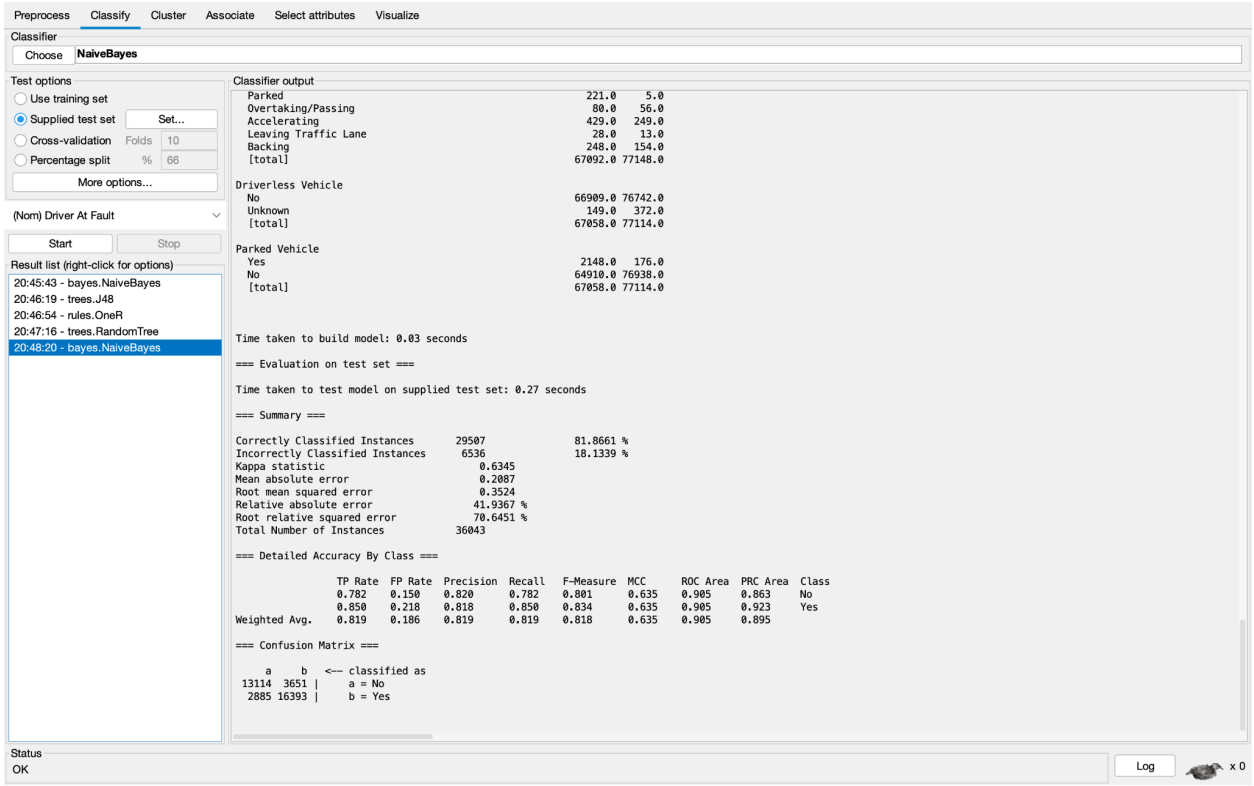
OneR: CorrelationAttributeEval



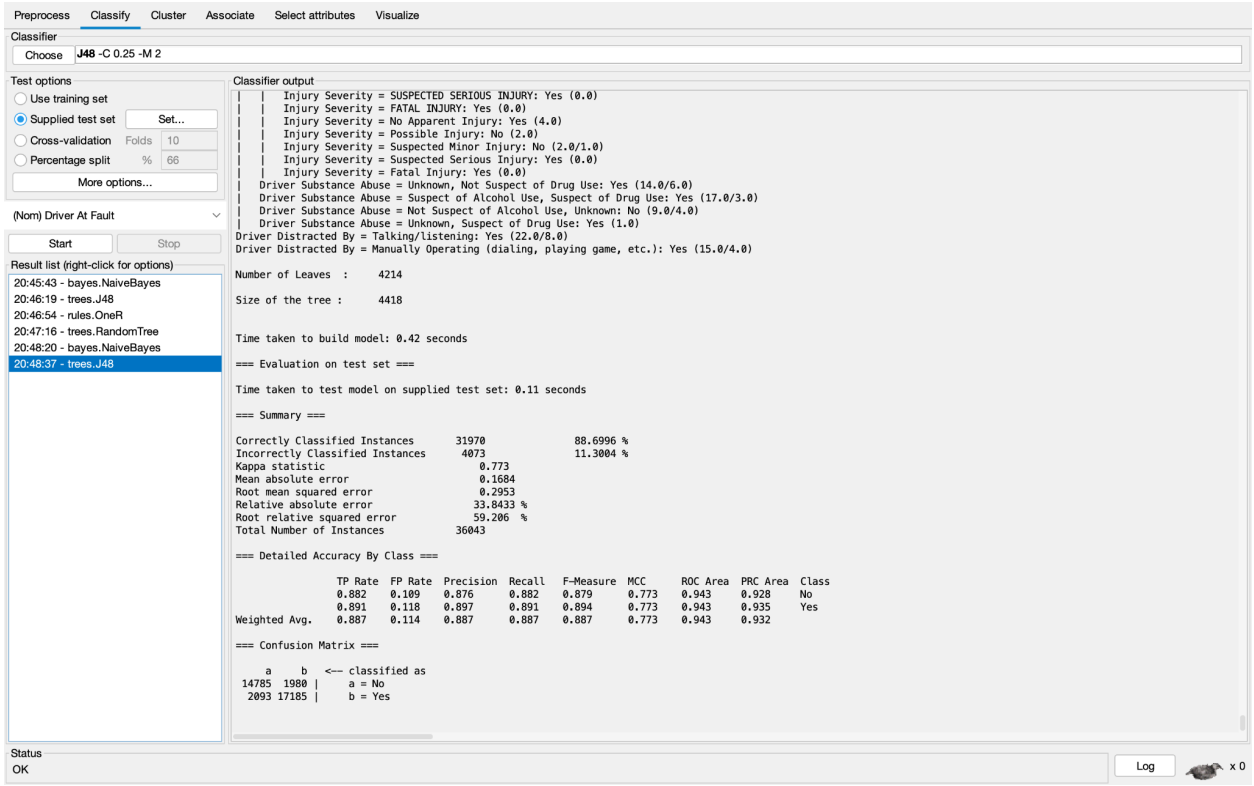
Random Tree: CorrelationAttributeEval



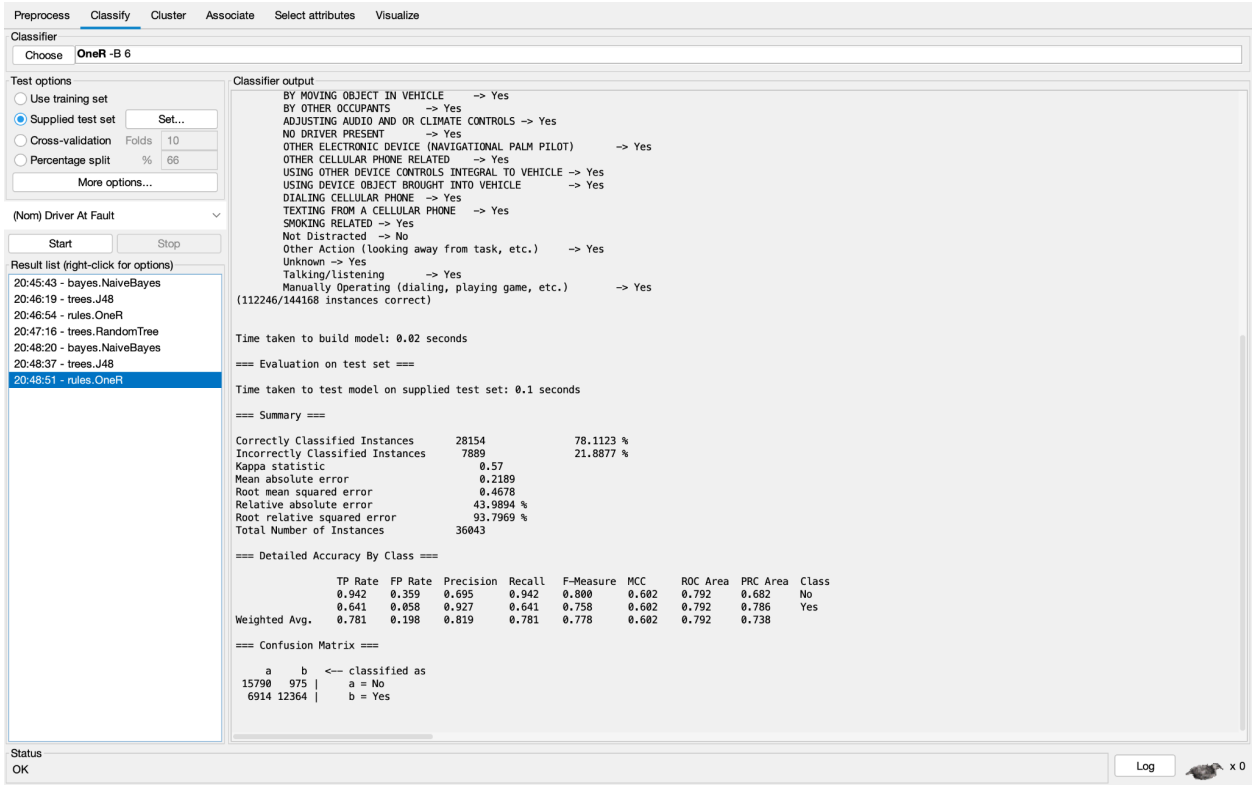
Naive Bayes: GainRatioEval



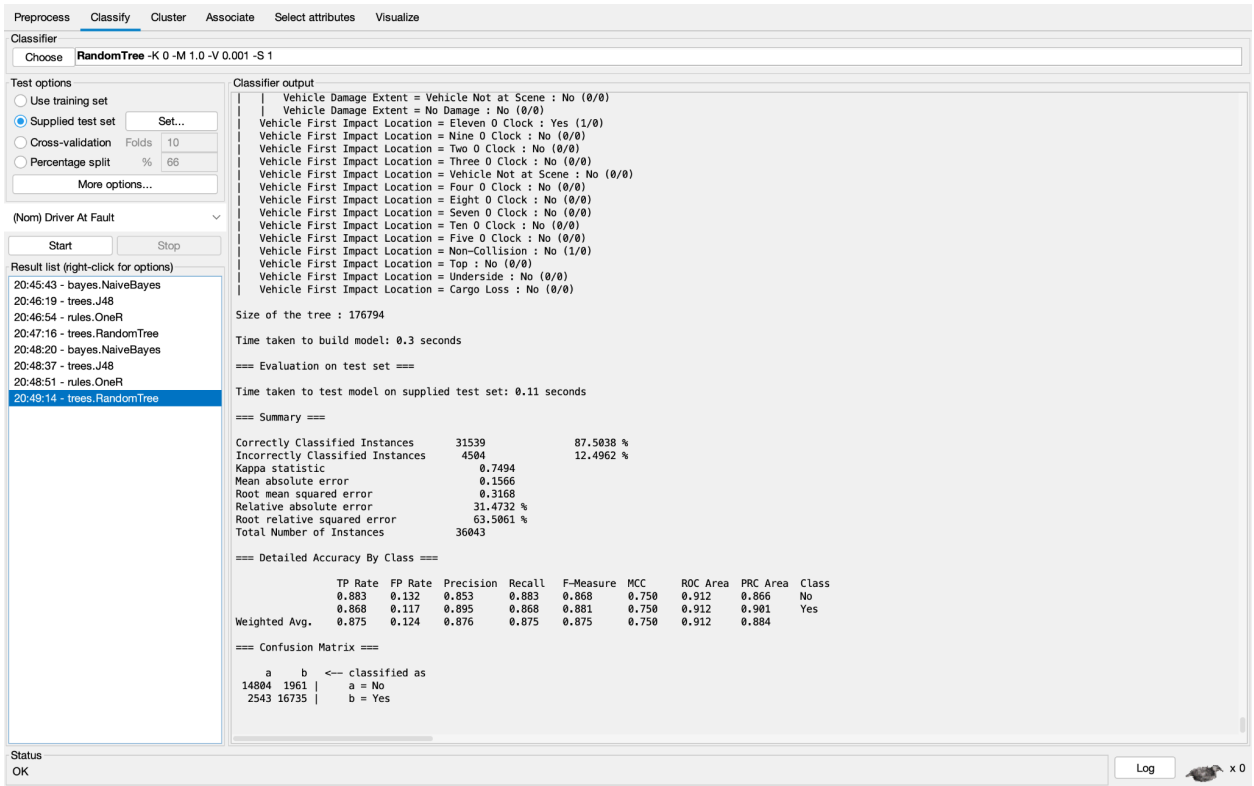
J48: GainRatioEval



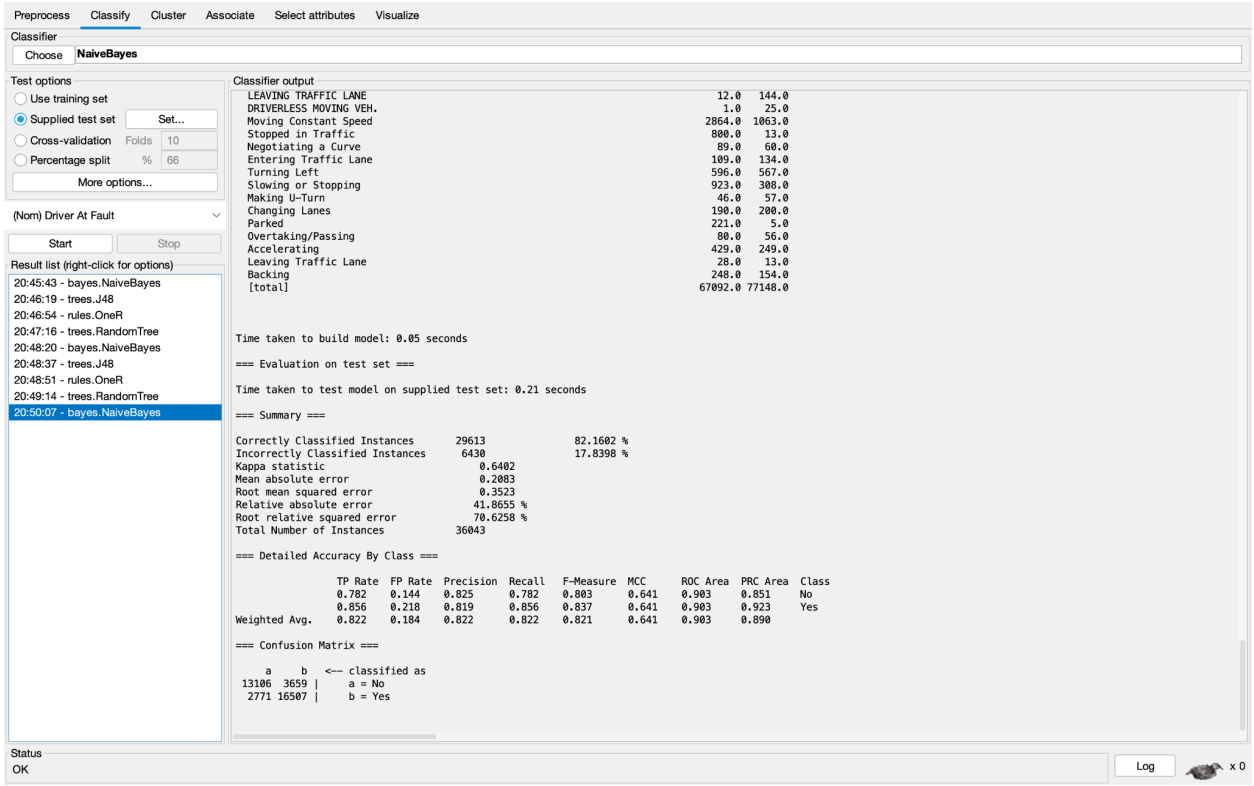
OneR: GainRatioEval



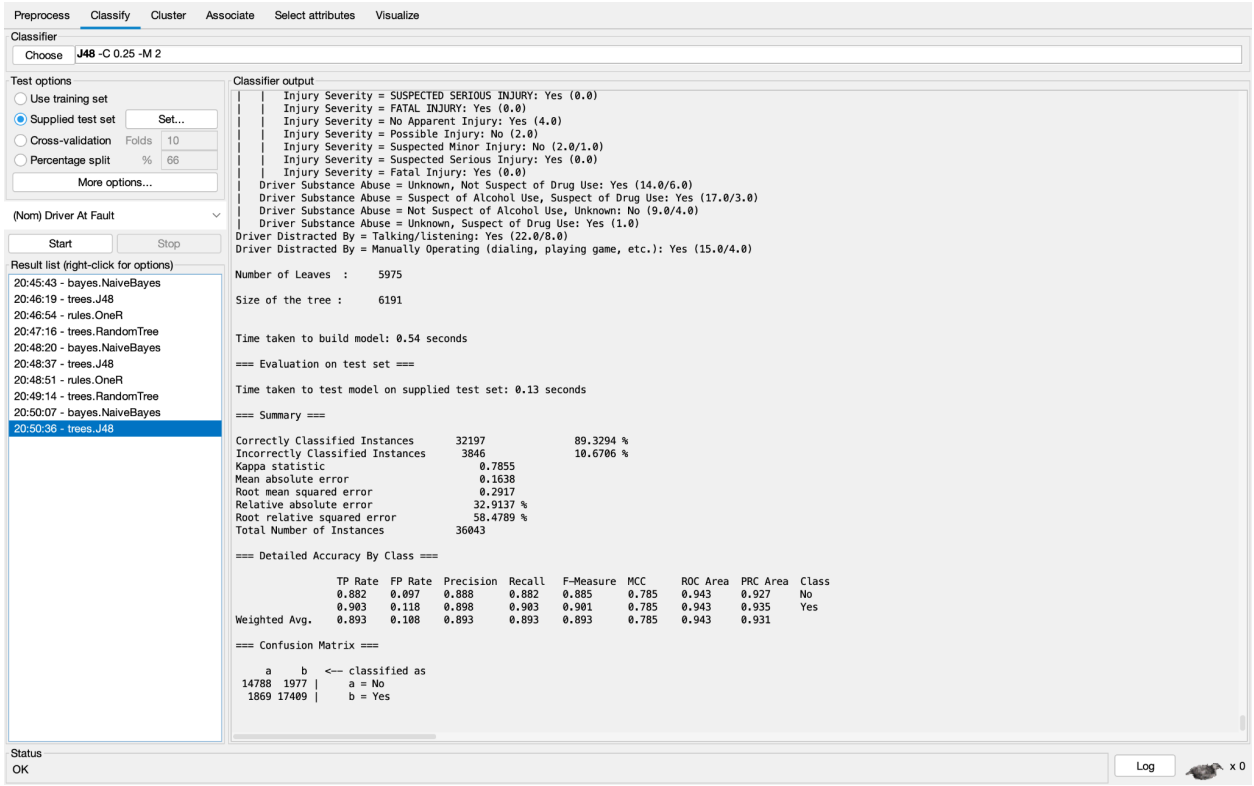
Random Tree: GainRatioEval



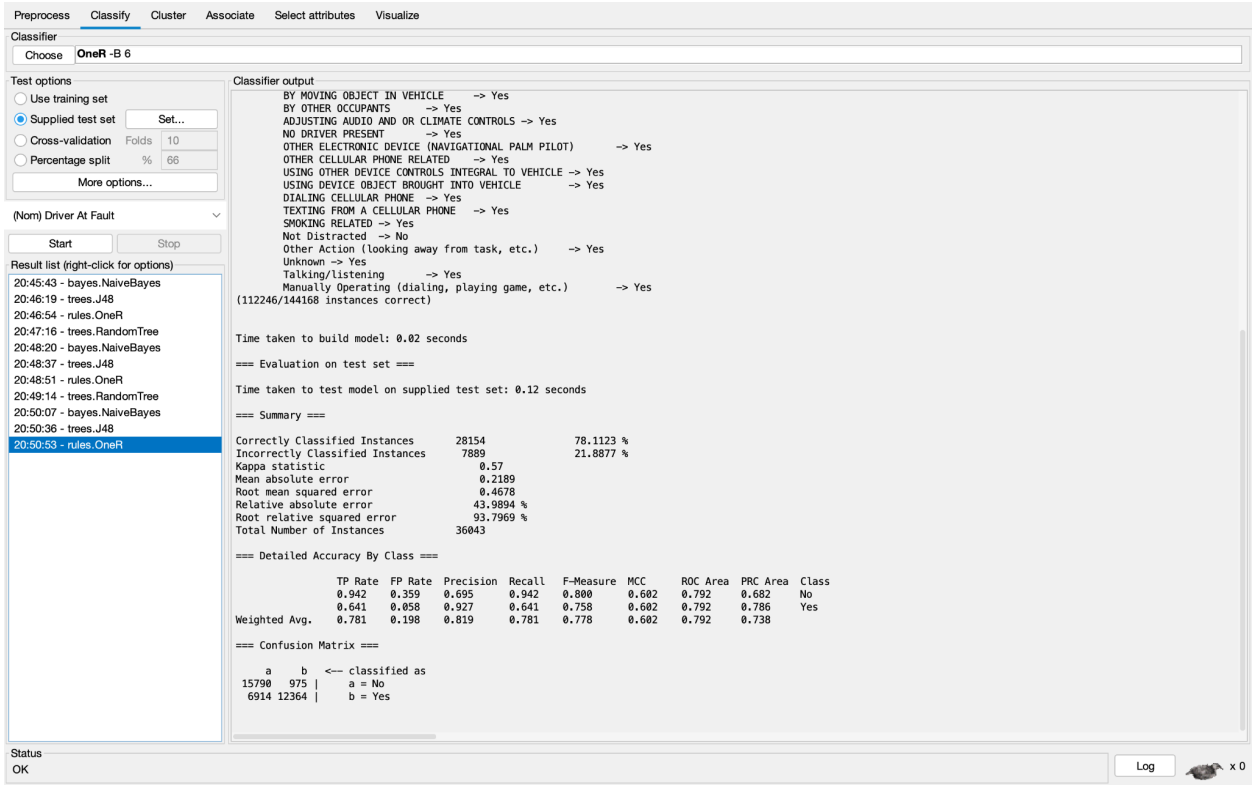
Naive Bayes: InfoGainEval



J48: InfoGainEval



OneR: InfoGainEval



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **OneR** -B 6

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Driver At Fault

Start Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR

Classifier output

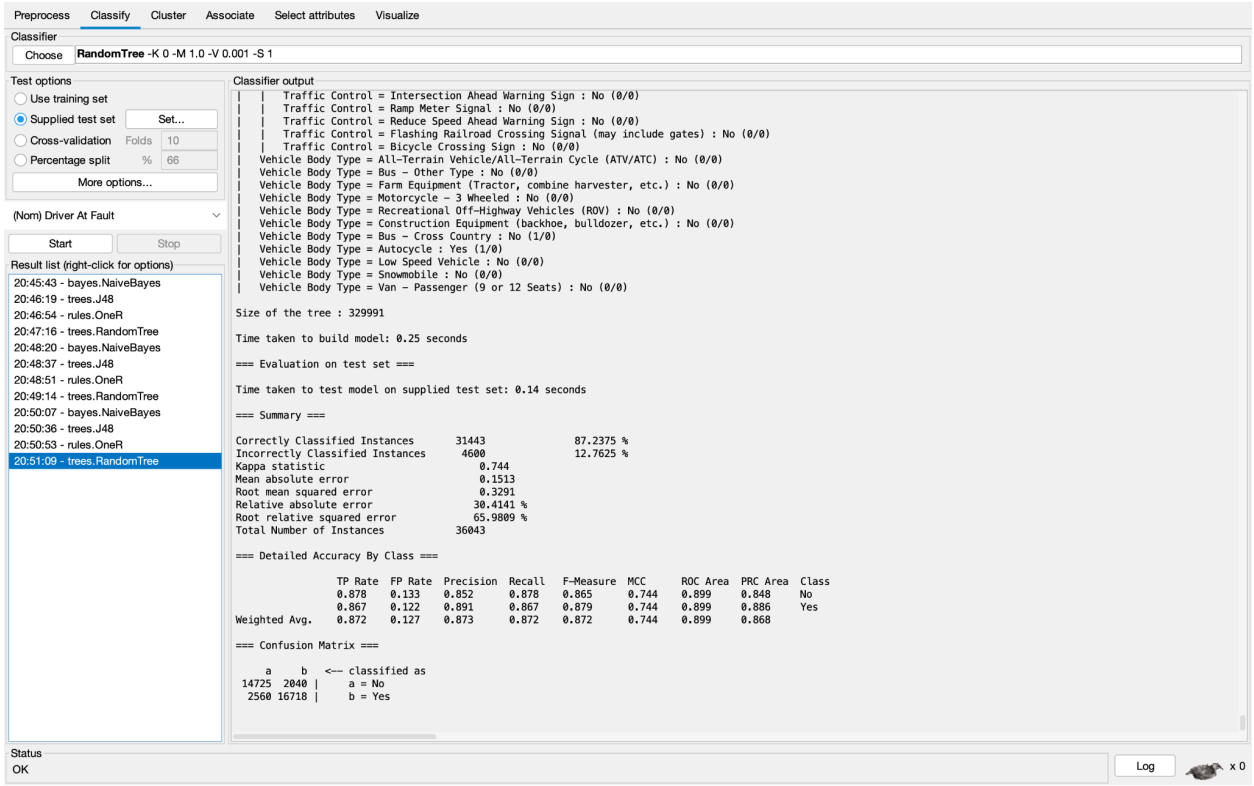
```
        BY MOVING OBJECT IN VEHICLE     -> Yes
        BY OTHER OCCUPANTS      -> Yes
        ADJUSTING AUDIO AND OR CLIMATE CONTROLS -> Yes
        NO DRIVER PRESENT       -> Yes
        OTHER ELECTRONIC DEVICE (NAVIGATIONAL PALM PILOT)       -> Yes
        OTHER CELLULAR PHONE RELATED    -> Yes
        USING OTHER DEVICE CONTROLS INTEGRAL TO VEHICLE -> Yes
        USING DEVICE OBJECT BROUGHT INTO VEHICLE        -> Yes
        DIALING CELLULAR PHONE  -> Yes
        TEXTING FROM A CELLULAR PHONE   -> Yes
        SMOKING RELATED -> Yes
        Not Distracted  -> No
        Other Action (looking away from task, etc.)     -> Yes
        Unknown -> Yes
        Talking/listening       -> Yes
        Manually Operating (dialing, playing game, etc.)        -> Yes
(112246/144168 instances correct)


Time taken to build model: 0.02 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.12 seconds

=== Summary ===

Correctly Classified Instances       28154               78.1123 %
Incorrectly Classified Instances      7889               21.8877 %
Kappa statistic                          0.57
Mean absolute error                      0.2189
Root mean squared error                  0.4678
Relative absolute error                 43.9894 %
Root relative squared error             93.7969 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.942    0.359    0.695      0.942    0.800      0.602    0.792     0.682     No
               0.641    0.058    0.927      0.641    0.758      0.602    0.792     0.786     Yes
Weighted Avg.  0.781    0.198    0.819      0.781    0.778      0.602    0.792     0.738

=== Confusion Matrix ===

     a     b   <-- classified as
 15790   975 |     a = No
  6914 12364 |     b = Yes
```

Status

OK

Log x 0

Random Tree: InfoGainEval



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **RandomTree** -K 0 -M 1.0 -V 0.001 -S 1

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Driver At Fault

Start Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR
20:51:09 - trees.RandomTree

Classifier output

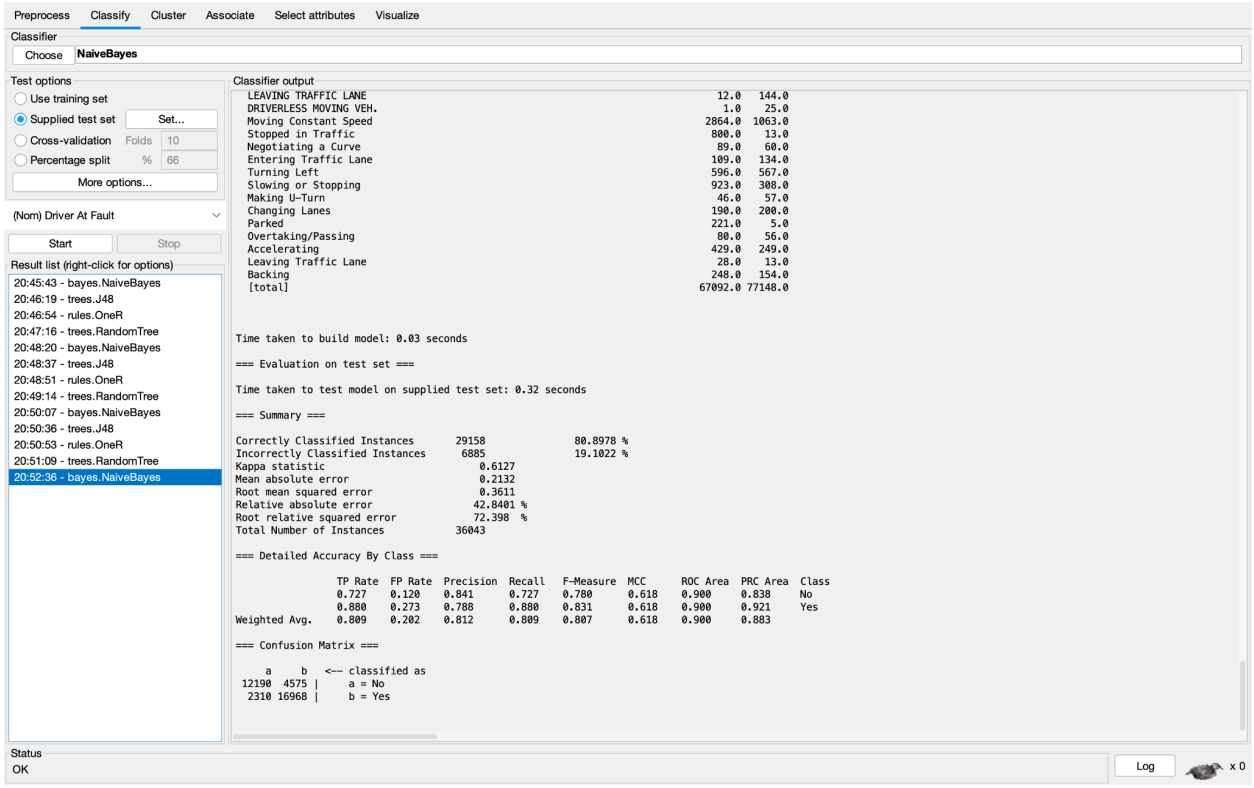
```
|   |   Traffic Control = Intersection Ahead Warning Sign : No (0/0)
|   |   Traffic Control = Ramp Meter Signal : No (0/0)
|   |   Traffic Control = Reduce Speed Ahead Warning Sign : No (0/0)
|   |   Traffic Control = Flashing Railroad Crossing Signal (may include gates) : No (0/0)
|   |   Traffic Control = Bicycle Crossing Sign : No (0/0)
|   Vehicle Body Type = All-Terrain Vehicle/All-Terrain Cycle (ATV/ATC) : No (0/0)
|   Vehicle Body Type = Bus - Other Type : No (0/0)
|   Vehicle Body Type = Farm Equipment (Tractor, combine harvester, etc.) : No (0/0)
|   Vehicle Body Type = Motorcycle - 3 Wheeled : No (0/0)
|   Vehicle Body Type = Recreational Off-Highway Vehicles (ROV) : No (0/0)
|   Vehicle Body Type = Construction Equipment (backhoe, bulldozer, etc.) : No (0/0)
|   Vehicle Body Type = Bus - Cross Country : No (1/0)
|   Vehicle Body Type = Autocycle : Yes (1/0)
|   Vehicle Body Type = Low Speed Vehicle : No (0/0)
|   Vehicle Body Type = Snowmobile : No (0/0)
|   Vehicle Body Type = Van - Passenger (9 or 12 Seats) : No (0/0)

Size of the tree : 329991

Time taken to build model: 0.25 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.14 seconds

=== Summary ===

Correctly Classified Instances       31443               87.2375 %
Incorrectly Classified Instances      4600               12.7625 %
Kappa statistic                          0.744
Mean absolute error                      0.1513
Root mean squared error                  0.3291
Relative absolute error                 30.4141 %
Root relative squared error             65.9809 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.878    0.133    0.852      0.878    0.865      0.744    0.899     0.848     No
               0.867    0.122    0.891      0.867    0.879      0.744    0.899     0.886     Yes
Weighted Avg.  0.872    0.127    0.873      0.872    0.872      0.744    0.899     0.868

=== Confusion Matrix ===

     a     b   <-- classified as
 14725  2040 |     a = No
  2560 16718 |     b = Yes
```

Status

OK

Log x 0

Naive Bayes: OneREval



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier
Choose **NaiveBayes**

Test options
- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66
- More options...

(Nom) Driver At Fault

Start | Stop

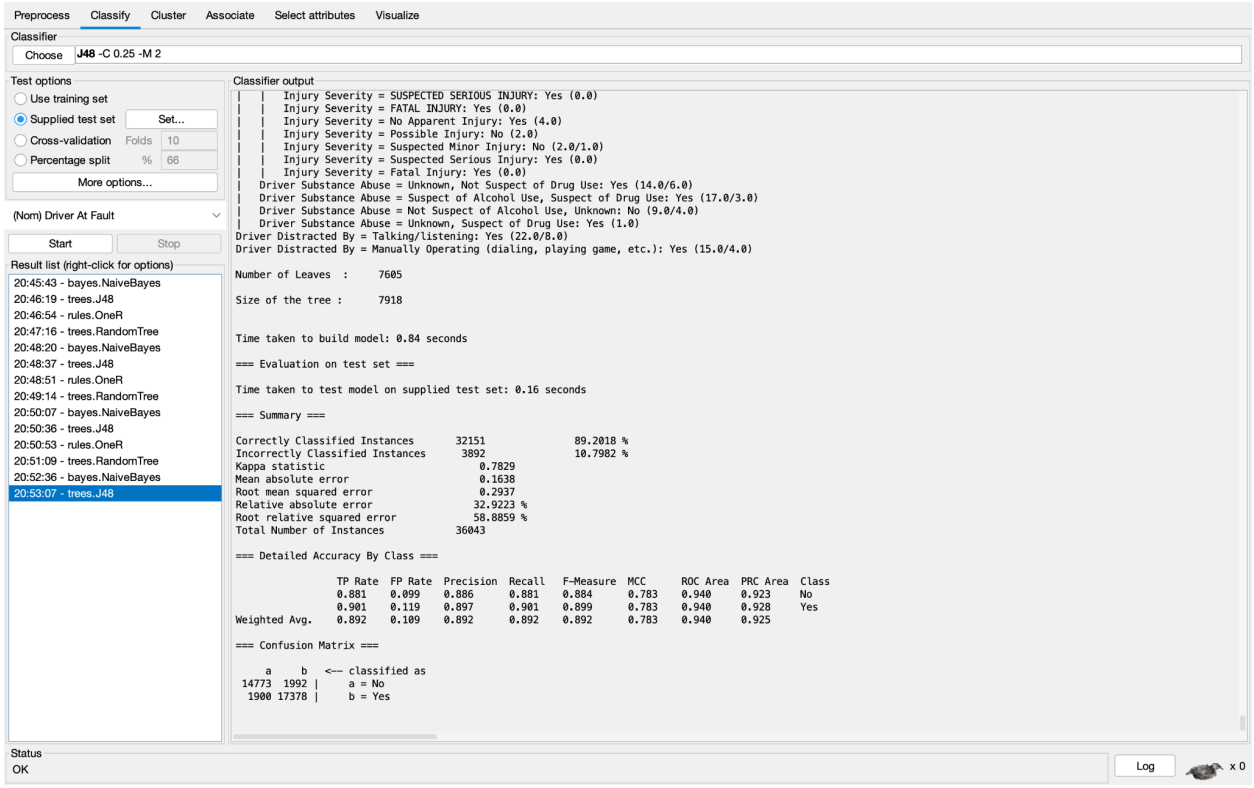
Result list (right-click for options)
- 20:45:43 - bayes.NaiveBayes
- 20:46:19 - trees.J48
- 20:46:54 - rules.OneR
- 20:47:16 - trees.RandomTree
- 20:48:20 - bayes.NaiveBayes
- 20:48:37 - trees.J48
- 20:48:51 - rules.OneR
- 20:49:14 - trees.RandomTree
- 20:50:07 - bayes.NaiveBayes
- 20:50:36 - trees.J48
- 20:50:53 - rules.OneR
- 20:51:09 - trees.RandomTree
- 20:52:36 - bayes.NaiveBayes

Classifier output

```
  LEAVING TRAFFIC LANE                                          12.0   144.0
  DRIVERLESS MOVING VEH.                                         1.0    25.0
  Moving Constant Speed                                       2864.0  1063.0
  Stopped in Traffic                                           800.0    13.0
  Negotiating a Curve                                           89.0    60.0
  Entering Traffic Lane                                        109.0   134.0
  Turning Left                                                 596.0   567.0
  Slowing or Stopping                                          923.0   308.0
  Making U-Turn                                                 46.0    57.0
  Changing Lanes                                               190.0   200.0
  Parked                                                       221.0     5.0
  Overtaking/Passing                                            80.0    56.0
  Accelerating                                                 429.0   249.0
  Leaving Traffic Lane                                          28.0    13.0
  Backing                                                      248.0   154.0
  [total]                                                    67092.0 77148.0


Time taken to build model: 0.03 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.32 seconds

=== Summary ===

Correctly Classified Instances       29158               80.8978 %
Incorrectly Classified Instances      6885               19.1022 %
Kappa statistic                          0.6127
Mean absolute error                      0.2132
Root mean squared error                  0.3611
Relative absolute error                 42.8401 %
Root relative squared error             72.398  %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.727    0.120    0.841      0.727    0.780      0.618    0.900     0.838     No
               0.880    0.273    0.788      0.880    0.831      0.618    0.900     0.921     Yes
Weighted Avg.  0.809    0.202    0.812      0.809    0.807      0.618    0.900     0.883

=== Confusion Matrix ===

     a     b   <-- classified as
 12190  4575 |     a = No
  2310 16968 |     b = Yes
```

Status
OK — Log — x 0

J48: OneREval



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier
Choose **J48** -C 0.25 -M 2

Test options
- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66
- More options...

(Nom) Driver At Fault

Start | Stop

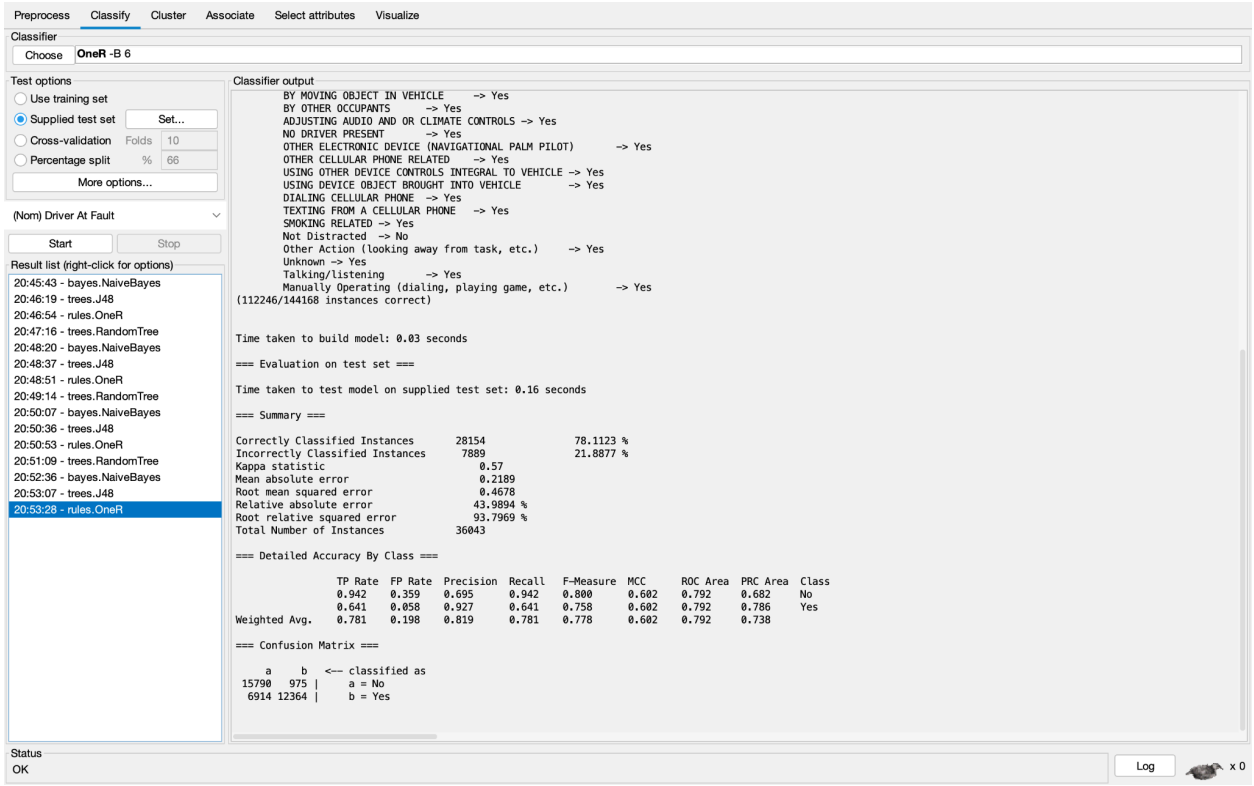
Result list (right-click for options)
- 20:45:43 - bayes.NaiveBayes
- 20:46:19 - trees.J48
- 20:46:54 - rules.OneR
- 20:47:16 - trees.RandomTree
- 20:48:20 - bayes.NaiveBayes
- 20:48:37 - trees.J48
- 20:48:51 - rules.OneR
- 20:49:14 - trees.RandomTree
- 20:50:07 - bayes.NaiveBayes
- 20:50:36 - trees.J48
- 20:50:53 - rules.OneR
- 20:51:09 - trees.RandomTree
- 20:52:36 - bayes.NaiveBayes
- 20:53:07 - trees.J48

Classifier output

```
|   |   Injury Severity = SUSPECTED SERIOUS INJURY: Yes (0.0)
|   |   Injury Severity = FATAL INJURY: Yes (0.0)
|   |   Injury Severity = No Apparent Injury: Yes (4.0)
|   |   Injury Severity = Possible Injury: No (2.0)
|   |   Injury Severity = Suspected Minor Injury: No (2.0/1.0)
|   |   Injury Severity = Suspected Serious Injury: Yes (0.0)
|   |   Injury Severity = Fatal Injury: Yes (0.0)
|   Driver Substance Abuse = Unknown, Not Suspect of Drug Use: Yes (14.0/6.0)
|   Driver Substance Abuse = Suspect of Alcohol Use, Suspect of Drug Use: Yes (17.0/3.0)
|   Driver Substance Abuse = Not Suspect of Alcohol Use, Unknown: No (9.0/4.0)
|   Driver Substance Abuse = Unknown, Suspect of Drug Use: Yes (1.0)
Driver Distracted By = Talking/listening: Yes (22.0/8.0)
Driver Distracted By = Manually Operating (dialing, playing game, etc.): Yes (15.0/4.0)

Number of Leaves  :     7605

Size of the tree :      7918


Time taken to build model: 0.84 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.16 seconds

=== Summary ===

Correctly Classified Instances       32151               89.2018 %
Incorrectly Classified Instances      3892               10.7982 %
Kappa statistic                          0.7829
Mean absolute error                      0.1638
Root mean squared error                  0.2937
Relative absolute error                 32.9223 %
Root relative squared error             58.8859 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.881    0.099    0.886      0.881    0.884      0.783    0.940     0.923     No
               0.901    0.119    0.897      0.901    0.899      0.783    0.940     0.928     Yes
Weighted Avg.  0.892    0.109    0.892      0.892    0.892      0.783    0.940     0.925

=== Confusion Matrix ===

     a     b   <-- classified as
 14773  1992 |     a = No
  1900 17378 |     b = Yes
```

Status
OK — Log — x 0

OneR: OneREval



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **OneR** -B 6

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Driver At Fault

Start Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR
20:51:09 - trees.RandomTree
20:52:36 - bayes.NaiveBayes
20:53:07 - trees.J48
20:53:28 - rules.OneR

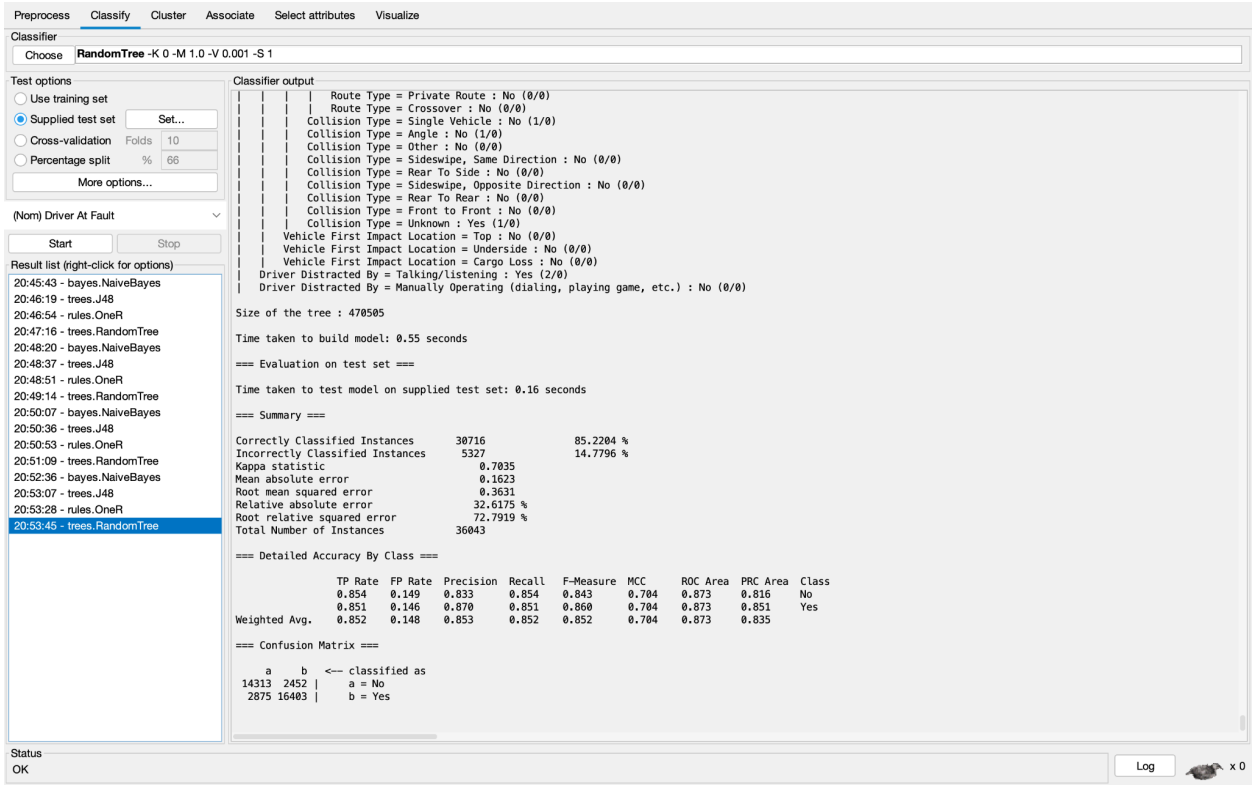
Classifier output

```
        BY MOVING OBJECT IN VEHICLE     -> Yes
        BY OTHER OCCUPANTS      -> Yes
        ADJUSTING AUDIO AND OR CLIMATE CONTROLS -> Yes
        NO DRIVER PRESENT       -> Yes
        OTHER ELECTRONIC DEVICE (NAVIGATIONAL PALM PILOT)       -> Yes
        OTHER CELLULAR PHONE RELATED    -> Yes
        USING OTHER DEVICE CONTROLS INTEGRAL TO VEHICLE -> Yes
        USING DEVICE OBJECT BROUGHT INTO VEHICLE        -> Yes
        DIALING CELLULAR PHONE  -> Yes
        TEXTING FROM A CELLULAR PHONE   -> Yes
        SMOKING RELATED -> Yes
        Not Distracted  -> No
        Other Action (looking away from task, etc.)     -> Yes
        Unknown -> Yes
        Talking/listening       -> Yes
        Manually Operating (dialing, playing game, etc.)        -> Yes
(112246/144168 instances correct)


Time taken to build model: 0.03 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.16 seconds

=== Summary ===

Correctly Classified Instances       28154               78.1123 %
Incorrectly Classified Instances      7889               21.8877 %
Kappa statistic                          0.57
Mean absolute error                      0.2189
Root mean squared error                  0.4678
Relative absolute error                 43.9894 %
Root relative squared error             93.7969 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.942    0.359    0.695      0.942    0.800      0.602    0.792     0.682     No
               0.641    0.058    0.927      0.641    0.758      0.602    0.792     0.786     Yes
Weighted Avg.  0.781    0.198    0.819      0.781    0.778      0.602    0.792     0.738

=== Confusion Matrix ===

     a     b   <-- classified as
 15790   975 |     a = No
  6914 12364 |     b = Yes
```

Status
OK

Log x 0

Random Tree: OneREval



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **RandomTree** -K 0 -M 1.0 -V 0.001 -S 1

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Driver At Fault

Start Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR
20:51:09 - trees.RandomTree
20:52:36 - bayes.NaiveBayes
20:53:07 - trees.J48
20:53:28 - rules.OneR
20:53:45 - trees.RandomTree

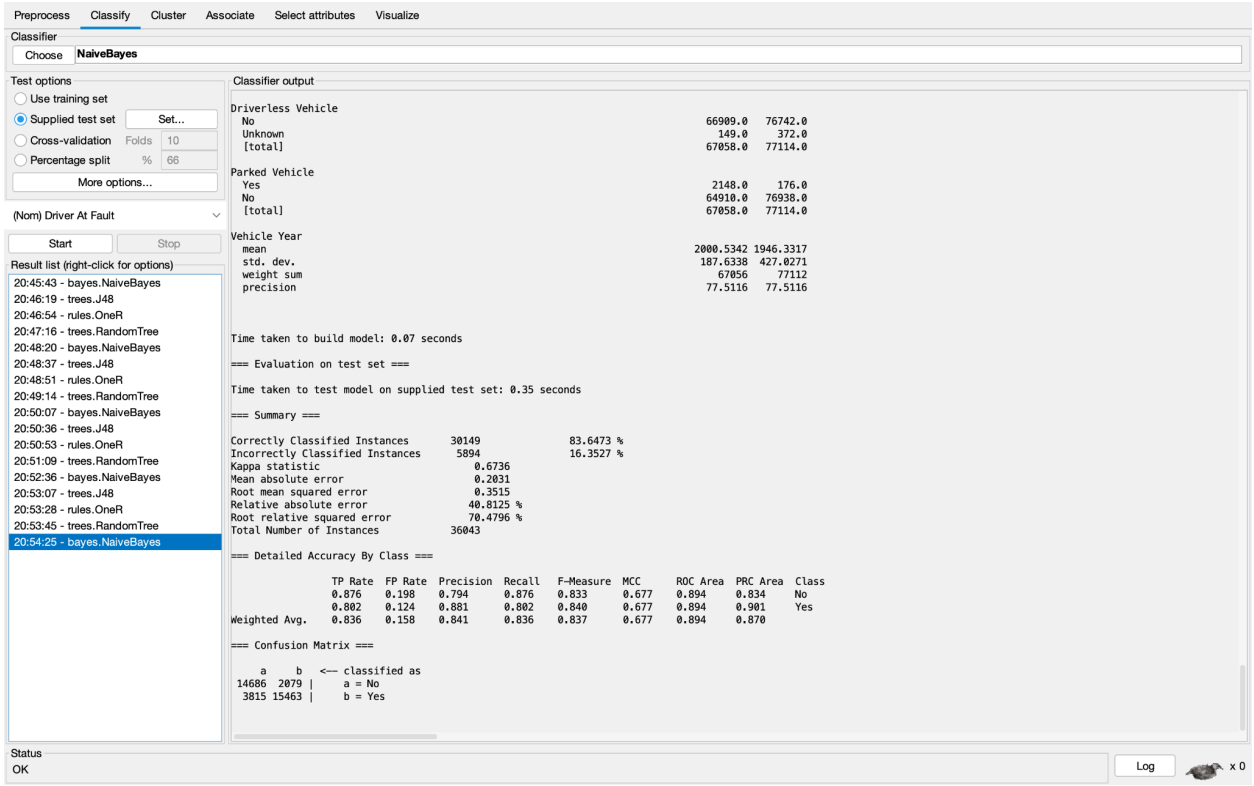
Classifier output

```
|   |   |   |   Route Type = Private Route : No (0/0)
|   |   |   |   Route Type = Crossover : No (0/0)
|   |   |   Collision Type = Single Vehicle : No (1/0)
|   |   |   Collision Type = Angle : No (1/0)
|   |   |   Collision Type = Other : No (0/0)
|   |   |   Collision Type = Sideswipe, Same Direction : No (0/0)
|   |   |   Collision Type = Rear To Side : No (0/0)
|   |   |   Collision Type = Sideswipe, Opposite Direction : No (0/0)
|   |   |   Collision Type = Rear To Rear : No (0/0)
|   |   |   Collision Type = Front to Front : No (0/0)
|   |   |   Collision Type = Unknown : Yes (1/0)
|   |   Vehicle First Impact Location = Top : No (0/0)
|   |   Vehicle First Impact Location = Underside : No (0/0)
|   |   Vehicle First Impact Location = Cargo Loss : No (0/0)
|   Driver Distracted By = Talking/listening : Yes (2/0)
|   Driver Distracted By = Manually Operating (dialing, playing game, etc.) : No (0/0)

Size of the tree : 470505

Time taken to build model: 0.55 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.16 seconds

=== Summary ===

Correctly Classified Instances       30716               85.2204 %
Incorrectly Classified Instances      5327               14.7796 %
Kappa statistic                          0.7035
Mean absolute error                      0.1623
Root mean squared error                  0.3631
Relative absolute error                 32.6175 %
Root relative squared error             72.7919 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.854    0.149    0.833      0.854    0.843      0.704    0.873     0.816     No
               0.851    0.146    0.870      0.851    0.860      0.704    0.873     0.851     Yes
Weighted Avg.  0.852    0.148    0.853      0.852    0.852      0.704    0.873     0.835

=== Confusion Matrix ===

     a     b   <-- classified as
 14313  2452 |     a = No
  2875 16403 |     b = Yes
```

Status
OK

Log x 0

Naive Bayes: Self-Selected



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **NaiveBayes**

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Driver At Fault

Start Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR
20:51:09 - trees.RandomTree
20:52:36 - bayes.NaiveBayes
20:53:07 - trees.J48
20:53:28 - rules.OneR
20:53:45 - trees.RandomTree
20:54:25 - bayes.NaiveBayes

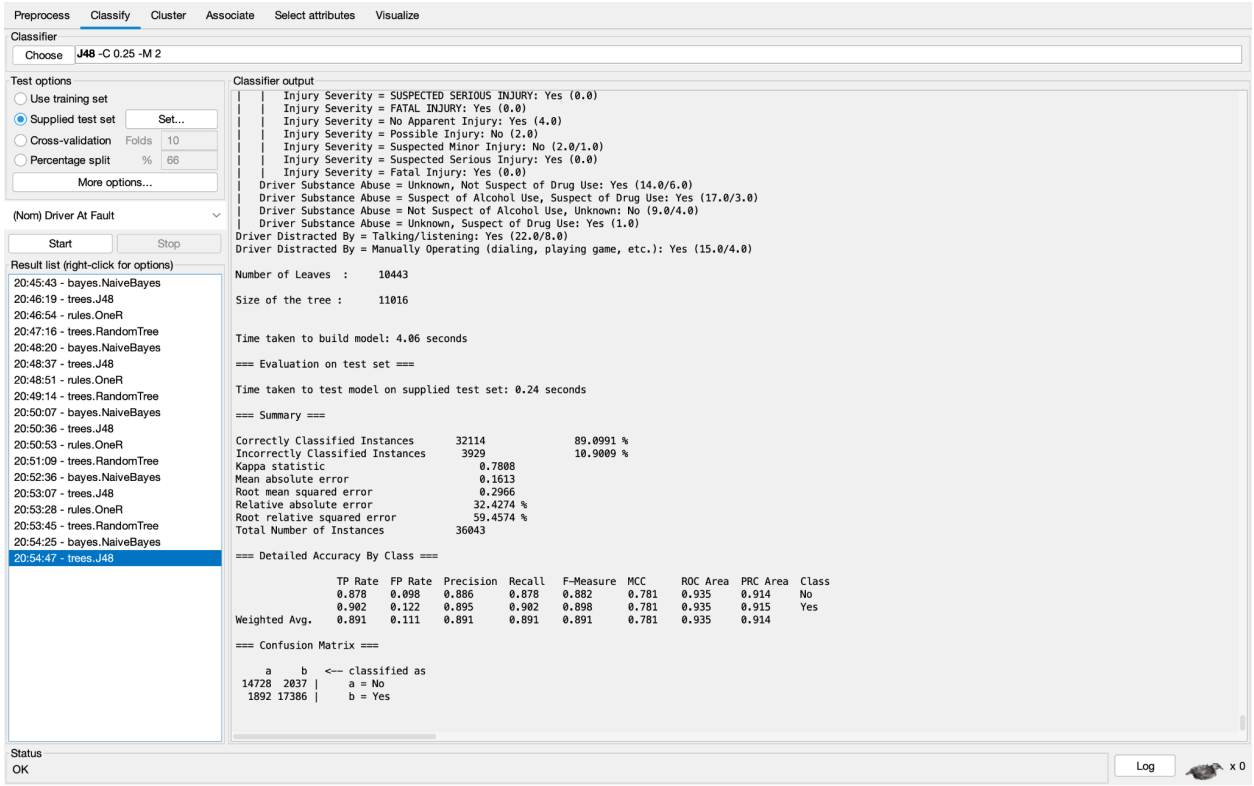
Classifier output

```
Driverless Vehicle
  No                                                          66909.0   76742.0
  Unknown                                                       149.0     372.0
  [total]                                                     67058.0   77114.0

Parked Vehicle
  Yes                                                          2148.0     176.0
  No                                                          64910.0   76938.0
  [total]                                                     67058.0   77114.0

Vehicle Year
  mean                                                      2000.5342 1946.3317
  std. dev.                                                  187.6338  427.0271
  weight sum                                                  67056     77112
  precision                                                   77.5116   77.5116


Time taken to build model: 0.07 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.35 seconds

=== Summary ===

Correctly Classified Instances       30149               83.6473 %
Incorrectly Classified Instances      5894               16.3527 %
Kappa statistic                          0.6736
Mean absolute error                      0.2031
Root mean squared error                  0.3515
Relative absolute error                 40.8125 %
Root relative squared error             70.4796 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.876    0.198    0.794      0.876    0.833      0.677    0.894     0.834     No
               0.802    0.124    0.881      0.802    0.840      0.677    0.894     0.901     Yes
Weighted Avg.  0.836    0.158    0.841      0.836    0.837      0.677    0.894     0.870

=== Confusion Matrix ===

     a     b   <-- classified as
 14686  2079 |     a = No
  3815 15463 |     b = Yes
```

Status
OK

Log x 0

J48: Self-Selected



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **J48** -C 0.25 -M 2

Test options

Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...

(Nom) Driver At Fault

Start Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR
20:51:09 - trees.RandomTree
20:52:36 - bayes.NaiveBayes
20:53:07 - trees.J48
20:53:28 - rules.OneR
20:53:45 - trees.RandomTree
20:54:25 - bayes.NaiveBayes
20:54:47 - trees.J48

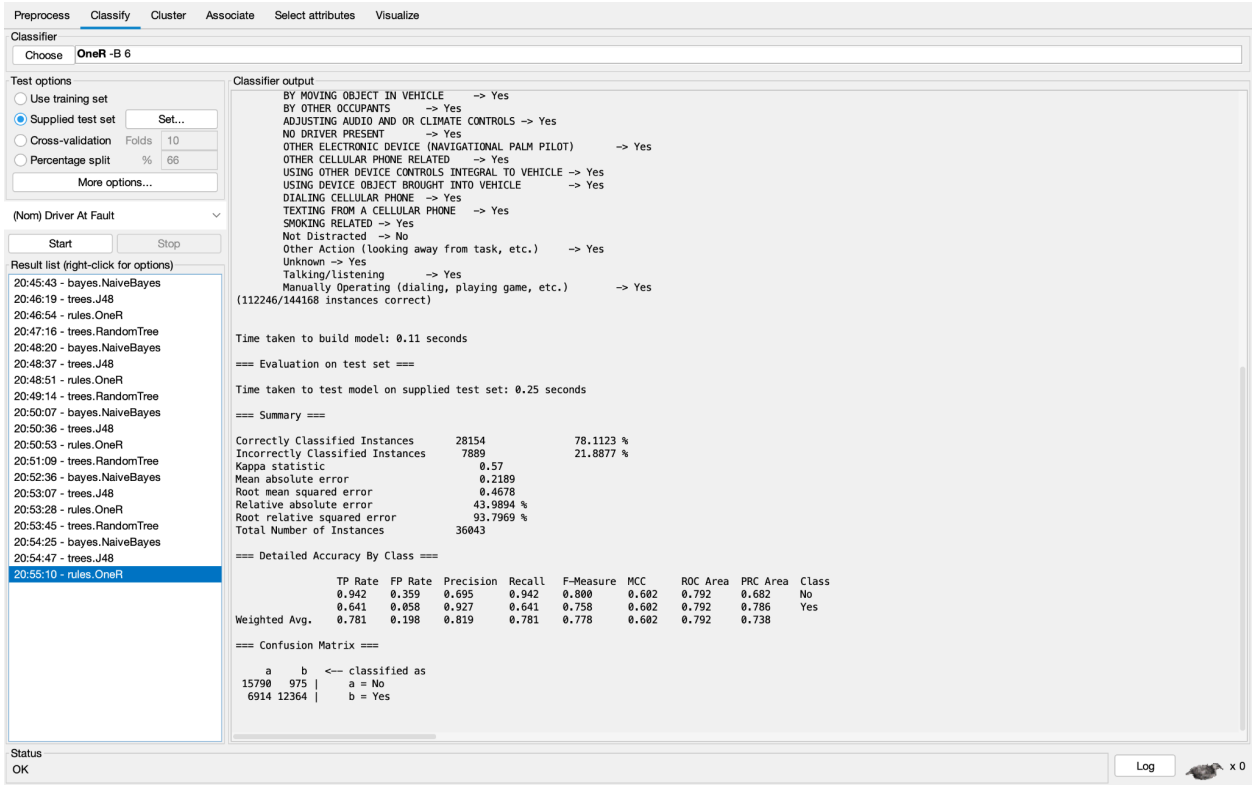
Classifier output

```
|   |   Injury Severity = SUSPECTED SERIOUS INJURY: Yes (0.0)
|   |   Injury Severity = FATAL INJURY: Yes (0.0)
|   |   Injury Severity = No Apparent Injury: Yes (4.0)
|   |   Injury Severity = Possible Injury: No (2.0)
|   |   Injury Severity = Suspected Minor Injury: No (2.0/1.0)
|   |   Injury Severity = Suspected Serious Injury: Yes (0.0)
|   |   Injury Severity = Fatal Injury: Yes (0.0)
|   Driver Substance Abuse = Unknown, Not Suspect of Drug Use: Yes (14.0/6.0)
|   Driver Substance Abuse = Suspect of Alcohol Use, Suspect of Drug Use: Yes (17.0/3.0)
|   Driver Substance Abuse = Not Suspect of Alcohol Use, Unknown: No (9.0/4.0)
|   Driver Substance Abuse = Unknown, Suspect of Drug Use: Yes (1.0)
Driver Distracted By = Talking/listening: Yes (22.0/8.0)
Driver Distracted By = Manually Operating (dialing, playing game, etc.): Yes (15.0/4.0)

Number of Leaves  :     10443

Size of the tree :      11016


Time taken to build model: 4.06 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.24 seconds

=== Summary ===

Correctly Classified Instances       32114               89.0991 %
Incorrectly Classified Instances      3929               10.9009 %
Kappa statistic                          0.7808
Mean absolute error                      0.1613
Root mean squared error                  0.2966
Relative absolute error                 32.4274 %
Root relative squared error             59.4574 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.878    0.098    0.886      0.878    0.882      0.781    0.935     0.914     No
               0.902    0.122    0.895      0.902    0.898      0.781    0.935     0.915     Yes
Weighted Avg.  0.891    0.111    0.891      0.891    0.891      0.781    0.935     0.914

=== Confusion Matrix ===

     a     b   <-- classified as
 14728  2037 |     a = No
  1892 17386 |     b = Yes
```

Status
OK

Log x 0

OneR: Self-Selected



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **OneR** -B 6

Test options

Use training set
Supplied test set — Set...
Cross-validation — Folds 10
Percentage split — % 66
More options...

(Nom) Driver At Fault

Start | Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR
20:51:09 - trees.RandomTree
20:52:36 - bayes.NaiveBayes
20:53:07 - trees.J48
20:53:28 - rules.OneR
20:53:45 - trees.RandomTree
20:54:25 - bayes.NaiveBayes
20:54:47 - trees.J48
20:55:10 - rules.OneR

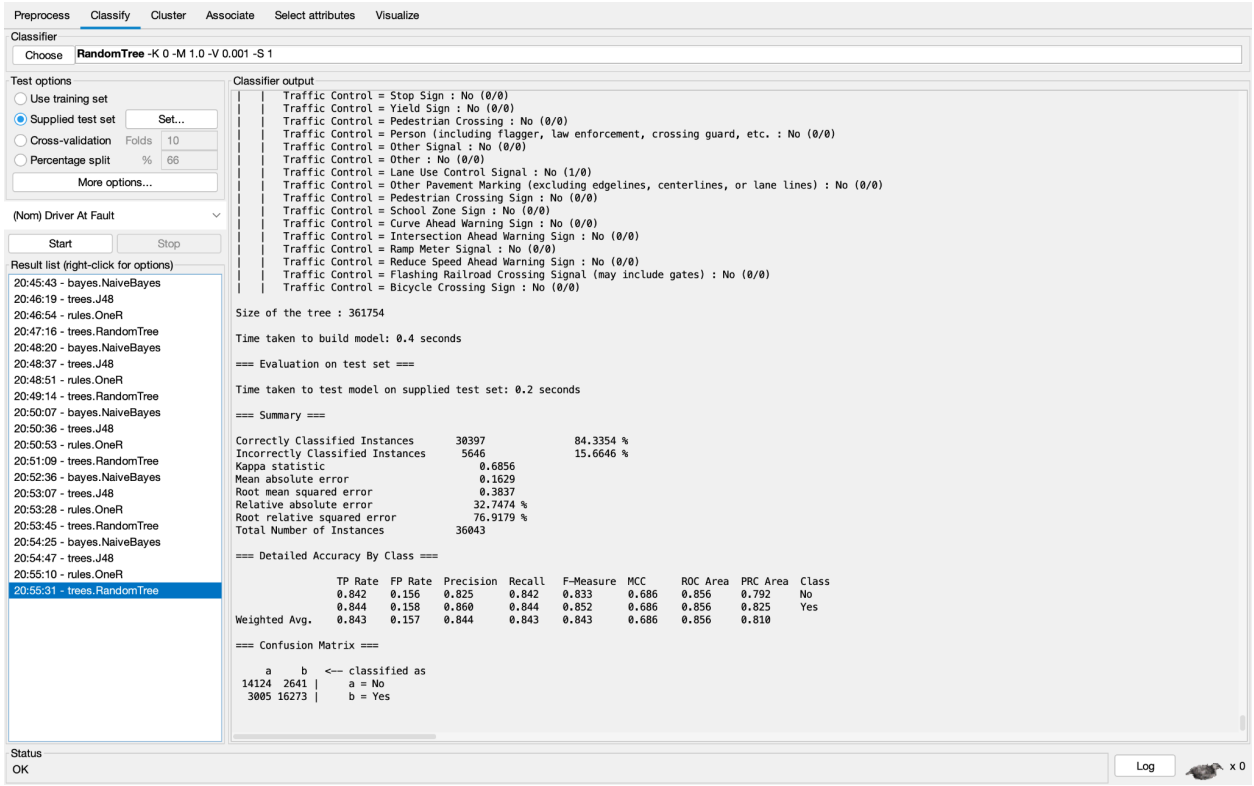
Classifier output

```
        BY MOVING OBJECT IN VEHICLE     -> Yes
        BY OTHER OCCUPANTS      -> Yes
        ADJUSTING AUDIO AND OR CLIMATE CONTROLS -> Yes
        NO DRIVER PRESENT       -> Yes
        OTHER ELECTRONIC DEVICE (NAVIGATIONAL PALM PILOT)       -> Yes
        OTHER CELLULAR PHONE RELATED    -> Yes
        USING OTHER DEVICE CONTROLS INTEGRAL TO VEHICLE -> Yes
        USING DEVICE OBJECT BROUGHT INTO VEHICLE        -> Yes
        DIALING CELLULAR PHONE  -> Yes
        TEXTING FROM A CELLULAR PHONE   -> Yes
        SMOKING RELATED -> Yes
        Not Distracted  -> No
        Other Action (looking away from task, etc.)     -> Yes
        Unknown -> Yes
        Talking/listening       -> Yes
        Manually Operating (dialing, playing game, etc.)        -> Yes
(112246/144168 instances correct)


Time taken to build model: 0.11 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.25 seconds

=== Summary ===

Correctly Classified Instances       28154               78.1123 %
Incorrectly Classified Instances      7889               21.8877 %
Kappa statistic                          0.57
Mean absolute error                      0.2189
Root mean squared error                  0.4678
Relative absolute error                 43.9894 %
Root relative squared error             93.7969 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.942    0.359    0.695      0.942    0.800      0.602    0.792     0.682     No
               0.641    0.058    0.927      0.641    0.758      0.602    0.792     0.786     Yes
Weighted Avg.  0.781    0.198    0.819      0.781    0.778      0.602    0.792     0.738

=== Confusion Matrix ===

     a     b   <-- classified as
 15790   975 |     a = No
  6914 12364 |     b = Yes
```

Status
OK

Log | x 0

Random Tree: Self-Selected



Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier

Choose **RandomTree** -K 0 -M 1.0 -V 0.001 -S 1

Test options

Use training set
Supplied test set — Set...
Cross-validation — Folds 10
Percentage split — % 66
More options...

(Nom) Driver At Fault

Start | Stop

Result list (right-click for options)

20:45:43 - bayes.NaiveBayes
20:46:19 - trees.J48
20:46:54 - rules.OneR
20:47:16 - trees.RandomTree
20:48:20 - bayes.NaiveBayes
20:48:37 - trees.J48
20:48:51 - rules.OneR
20:49:14 - trees.RandomTree
20:50:07 - bayes.NaiveBayes
20:50:36 - trees.J48
20:50:53 - rules.OneR
20:51:09 - trees.RandomTree
20:52:36 - bayes.NaiveBayes
20:53:07 - trees.J48
20:53:28 - rules.OneR
20:53:45 - trees.RandomTree
20:54:25 - bayes.NaiveBayes
20:54:47 - trees.J48
20:55:10 - rules.OneR
20:55:31 - trees.RandomTree

Classifier output

```
|   |   Traffic Control = Stop Sign : No (0/0)
|   |   Traffic Control = Yield Sign : No (0/0)
|   |   Traffic Control = Pedestrian Crossing : No (0/0)
|   |   Traffic Control = Person (including flagger, law enforcement, crossing guard, etc. : No (0/0)
|   |   Traffic Control = Other Signal : No (0/0)
|   |   Traffic Control = Other : No (0/0)
|   |   Traffic Control = Lane Use Control Signal : No (1/0)
|   |   Traffic Control = Other Pavement Marking (excluding edgelines, centerlines, or lane lines) : No (0/0)
|   |   Traffic Control = Pedestrian Crossing Sign : No (0/0)
|   |   Traffic Control = School Zone Sign : No (0/0)
|   |   Traffic Control = Curve Ahead Warning Sign : No (0/0)
|   |   Traffic Control = Intersection Ahead Warning Sign : No (0/0)
|   |   Traffic Control = Ramp Meter Signal : No (0/0)
|   |   Traffic Control = Reduce Speed Ahead Warning Sign : No (0/0)
|   |   Traffic Control = Flashing Railroad Crossing Signal (may include gates) : No (0/0)
|   |   Traffic Control = Bicycle Crossing Sign : No (0/0)

Size of the tree : 361754

Time taken to build model: 0.4 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.2 seconds

=== Summary ===

Correctly Classified Instances       30397               84.3354 %
Incorrectly Classified Instances      5646               15.6646 %
Kappa statistic                          0.6856
Mean absolute error                      0.1629
Root mean squared error                  0.3837
Relative absolute error                 32.7474 %
Root relative squared error             76.9179 %
Total Number of Instances            36043

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.842    0.156    0.825      0.842    0.833      0.686    0.856     0.792     No
               0.844    0.158    0.860      0.844    0.852      0.686    0.856     0.825     Yes
Weighted Avg.  0.843    0.157    0.844      0.843    0.843      0.686    0.856     0.810

=== Confusion Matrix ===

     a     b   <-- classified as
 14124  2641 |     a = No
  3005 16273 |     b = Yes
```

Status
OK

Log | x 0

Performance Metrics Used

We determined that accuracy would be a good measure of performance for our models because our data had very little skew (56% yes, and 44% no) with only two classes. Additionally, we claim that other metrics of performance such as recall which are based on False Negative rates would not be very useful because the classes “Yes” and “No” are essentially symmetric. (For any use case of such a model, if the driver is not at fault, then the other party involved must be at fault. Thus, the consequences of false-negatives and false-positives are identical, so accuracy is the only important metric).

The J48 model with the InfoGainEval dataset had the best accuracy of 89%.
Highest TP Rate: 89.3%
Highest TN Rate:
Lowest Mean Squared Error: 0.264
This accuracy is decently high and provides decent prediction of who is at fault

How to Reproduce

1. Under the Preprocess tab, remove the attributes which prevent arff to csv conversion
2. Open Weka and load the “rawdata.csv” dataset.
3. Remove redundant attributes
4. Parse dates into only the minutes using the python code (code included on report)
5. Select the attributes according to the corresponding attribute selection algorithm.
6. Split the data into train and test splits.
7. Train the desired model (J48) using the test set as a “supplied test set”.
8. Click Start.