

Predicting Driver at Fault in a Vehicle Collision

Julia Chen and Yuan Yin

Thomas Jefferson High School for Science and Technology

Machine Learning 1

Dr. Yilmaz

October 23, 2024

Introduction

We chose a dataset on data.gov, which is an official website of the United States. The link to the dataset is: <https://catalog.data.gov/dataset/crash-reporting-drivers-data>. The dataset contains information on motor vehicle operators involved in traffic collisions within Montgomery County in Maryland. The data was collected with the Automated Crash Reporting System (ACRS) of the Maryland State Police and is updated weekly on the website. We downloaded the dataset on 9/19/24.

The model that we create will be useful in situations where there are two drivers involved in a vehicle collision and do not agree on who is at fault. Based on the circumstances of the crash, our model can determine which driver is at fault.

Description of Dataset

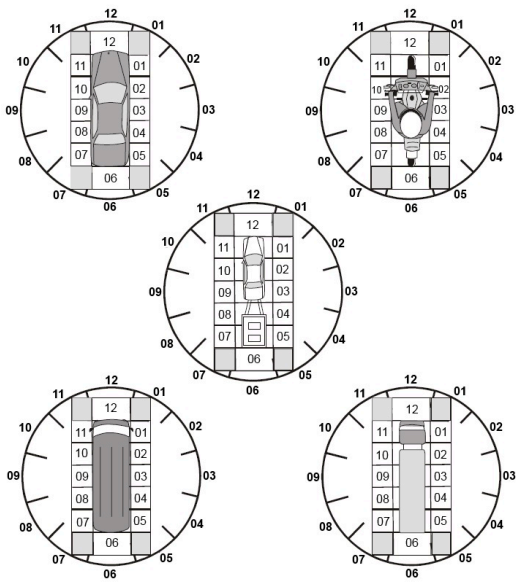
This dataset contains 39 attributes (before removal) and 184897 instances of vehicle crashes. The dimension of the dataset is 39.

Table 1 contains information about the attributes and their missing values.

Table 1

Attribute Name	Description of Attribute	Number of Missing Values	Percentage of Missing Values
Report Number	The report number of the report of the crash	0	0%
Local Case Number	Identification number of the local case of the crash	0	0%
Agency Name	Name of police department that investigated the crash	0	0%
ACRS Report Type	The type of crash as categorized by the Maryland State Police's Automated Crash Reporting System (Fatal Crash, Injury Crash, etc.)	0	0%
Crash Date/Time	The date and time the crash occurred	0	0%

Route Type	The type of road that the crash happened on	18161	9.822%
Road Name	The name of the road that the crash happened on	18708	10.118%
Cross-Street Name	The name of the street that was intersecting the road where the crash happened	23871	12.910%
Off-Road Description	The name of the place that the crash happened at if it was not on the road (parking lots)	167805	90.756%
Municipality	Name of the municipality the crash took place in if it happened in a municipality	165771	89.656%
Related Non-Motorist	Type of non-motorist that was related to the crash if there was one (pedestrian, cyclist, etc.)	178999	96.810%
Collision Type	The type of collision that happened (Front to Rear, Head On, etc.)	1388	0.751%
Weather	Weather conditions at the time and place of crash	14133	7.644%
Surface Condition	Condition of the surface that the crash happened on (snow, ice, etc.)	21848	11.816%
Light	Light conditions of the time and place of the crash (Daylight, Dawn, etc.)	2165	1.171%
Traffic Control	Name of traffic sign that was controlling the traffic around the location of the crash	26879	14.537%
Driver Substance Abuse	Substances that the driver was suspected or not suspected of using	44665	24.156%
Non-Motorist Substance Abuse	Substances that the non-motorist was using or not using	180366	97.549%
Person ID	Driver identification number	0	0%
Injury Severity	The severity of the injury as a result of the crash (fatal injury, possible injury, etc.)	789	0.427%

Circumstance	Road conditions that contributed to the crash (wet, icy, road construction, etc.)	150045	81.151%
Driver Distracted By	What the driver was distracted by, if they were distracted	36683	19.840%
Drivers License State	Abbreviation of the state that issued the driver's drivers license	11432	6.183%
Vehicle ID	Vehicle identification number	0	0%
Vehicle Damage Extent	Amount of damage sustained by the vehicle (no damage, superficial, functional, disabling, destroyed)	6936	3.751%
Vehicle First Impact Location	<p>Area of the vehicle that received the initial impact (reported as a clockpoint: one o'clock, two o'clock, etc.)</p> <p>Figure 1</p> <p>CLOCKPOINT DIAGRAM</p>  <p><i>Note. From TraCS, Clockpoint diagram</i></p>	3238	1.751%
Vehicle Body Type	Configuration of the vehicle (passenger car, van, bus, utility vehicle, etc.)	3903	2.111%

Vehicle Movement	How the vehicle was moving when the crash occurred (moving constant speed, accelerating, backing, parking, etc.)	4133	2.235%
Vehicle Going Dir	Direction the vehicle was traveling when the crash occurred (north/south/east/west)	10108	5.467%
Speed Limit	Speed limit at the location of the crash	5909	3.196%
Driverless Vehicle	Was the vehicle driverless when the crash occurred? (Binary: yes/no)	741	0.401%
Parked Vehicle	Was the vehicle parked when the crash occurred? (Binary: yes/no)	1534	0.830%
Vehicle Year	Year the vehicle was manufactured	4340	2.347%
Vehicle Make	Model of the vehicle involved in the crash	4089	2.212%
Vehicle Model	Model of the vehicle involved in the crash	4722	2.554%
Latitude	Latitude of crash	0	0%
Longitude	Latitude of crash	0	0%
Location	Location of crash (in the form (Longitude, Latitude))	0	0%
Driver At Fault (Chosen Class)	Is the driver described in the crash report responsible for the crash? (Binary: yes/no)	4686	2.534%

We have chosen to predict the class “Driver at Fault,” a binary attribute stating whether the driver described in the report is at fault in a vehicle crash. The possible values for this class are “Yes” and “No”. The values for our chosen class are slightly skewed towards “Yes”; there are more instances of “Yes” than “No” in the “Driver at Fault” column of the dataset. The class distribution for this dataset are: 96390 instances of “Yes” (52.132%) and 83821 instances of “No” (45.334%).

Preprocessing

Remove Instances Missing the Class

Instances that are missing the class value, “Driver at Fault,” must be removed since we cannot replace missing values in the class attribute. The number of instances in the dataset was reduced from 184897 to 180211.

Remove Attributes Missing More Than 70% of Their Values

Attributes that have a large amount of missing values will not be useful for predicting the class because there is not enough information to accurately replace all of the missing values. Thus, we removed attributes that are missing more than 70% of their values: Off-Road Description, Municipality, Related Non-Motorist, Non-Motorist Substance Abuse, and Circumstance. This reduced the number of attributes in the dataset to 33.

Remove Derived Attributes

To reduce redundancy, we removed derived values. The Location attribute is derived from the Longitude and Latitude attributes, so we removed it, bringing the total number of attributes in the dataset to 32.

Remove Unnecessary Attributes

Attributes that were deemed not useful to the dataset were removed. We removed the attribute Driverless Vehicle as all of the values were “No” except for less than 0.5% of values, which were unknown. Thus, this attribute is extremely unlikely to have a relationship with the class. For attributes Person ID and Vehicle ID, all values were distinct and random, so it is likely that these attributes would not have a relationship with the class either. We removed these attributes as well.

Stratified Random Sampling

We used StratifiedRemoveFolds (filters → supervised → instance) with a numFolds of 18 to get a sample size of 10012 with preserved proportions in the class. As a result, we have 4657 values of “No” and 5355 values of “Yes” in the class.

Unify Values

We used MergeManyValues to combine values within an attribute that represented the same value. For example, in the Weather attribute, there were values of “CLEAR” as well as “Clear,” which represents the same value. We first needed to convert all String values into Nominal, which we only needed to do for our Local Case Number attribute. We used the filter StringToNominal (filters → unsupervised → attribute). Then, we wrote a script to turn every alphabetic character to uppercase, which combined values that were the same but had different casing. For other cases, such as combining “TOP” and “ROOF TOP” in the Vehicle First Impact

Location, we combined the two values into one using MergeManyValues (filters → unsupervised → attribute). The attributes that contained values that needed to be unified were: Agency Name, Route Type, Weather, Surface Condition, Light, Traffic Control, Driver Substance Abuse, Injury Severity, Driver Distracted By, Vehicle Damage Extent, Vehicle First Impact Location, Vehicle Body Type, and Vehicle Movement.

Fill in Missing Values

We filled in the missing values within the attributes. We used the mean for numeric values and the mode for qualitative values to fill in missing values.

Attribute Selection

CorrelationAttributeEval

CorrelationAttributeEval is an attribute selection algorithm that evaluates the correlation between an attribute and the class variable by calculating the Pearson Correlation Coefficient, shown below.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

The Pearson Correlation Coefficient ranges from [-1, 1], with -1 indicating a strong negative correlation and 1 indicating a strong positive correlation. A greater absolute value of the Pearson Correlation Coefficient corresponds to an attribute that is strongly correlated with the class. A Pearson Correlation Coefficient value that is close to 0 indicates that a value has a weak correlation with the class. CorrelationAttributeEval ranks attributes that are strongly correlated with the class higher than attributes that have a weak correlation with the class.

The screenshot shows the CorrelationAttributeEval interface. The 'Attribute Selection Mode' is set to 'Use full training set'. The 'Attribute selection output' shows a list of ranked attributes. The top attributes are:

Rank	Attribute
0.3668	16 Driver Distracted By
0.2089	18 Vehicle First Impact Location
0.1219	19 Vehicle Damage Extent
0.1137	23 Speed Limit
0.1097	24 Parked Vehicle
0.1029	15 Injury Severity
0.1007	21 Vehicle Movement
0.0727	13 Traffic Control
0.0616	9 Collision Type
0.0538	25 Vehicle Year
0.0505	4 Abuse Report Type
0.0349	12 Light
0.0263	28 Latitude
0.0237	6 Route Type
0.0207	7 Road Name
0.0201	3 Agency Name
0.0197	14 Driver Substance Abuse
0.0189	20 Vehicle Body Type
0.0184	29 Longitude
0.0164	22 Vehicle Going Dir
0.0125	8 Cross-Street Name
0.0108	27 Vehicle Model

The screenshot shows the CorrelationAttributeEval interface with a different set of attributes. The 'Attribute Selection Mode' is set to 'Use full training set'. The 'Attribute selection output' shows a list of ranked attributes. The top attributes are:

Rank	Attribute
0.1029	15 Injury Severity
0.1007	21 Vehicle Movement
0.0727	13 Traffic Control
0.0616	9 Collision Type
0.0538	25 Vehicle Year
0.0505	4 Abuse Report Type
0.0349	12 Light
0.0263	28 Latitude
0.0237	6 Route Type
0.0207	7 Road Name
0.0201	3 Agency Name
0.0197	14 Driver Substance Abuse
0.0189	20 Vehicle Body Type
0.0184	29 Longitude
0.0164	22 Vehicle Going Dir
0.0125	8 Cross-Street Name
0.0108	27 Vehicle Model
0.0097	26 Vehicle Make
0.0095	1 Report Number
0.0095	2 Local Case Number
0.0095	5 Crash Date/Time
0.0051	17 Drivers License State
0.0045	11 Surface Condition
0.0042	10 Weather

Selected attributes: 16, 19, 18, 23, 24, 15, 21, 13, 9, 25, 4, 12, 28, 6, 7, 3, 14, 20, 29, 22, 8, 27, 26, 1, 2, 5, 17, 11, 10 : 29

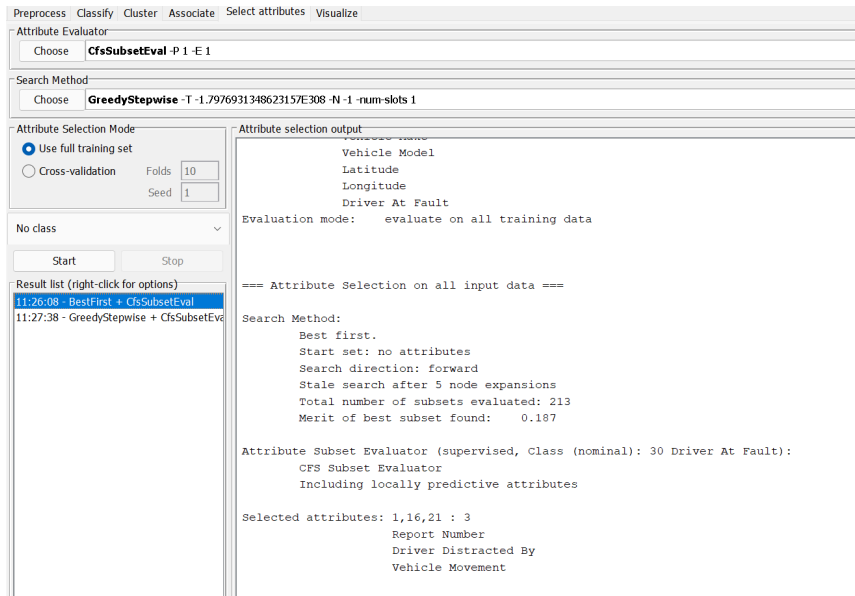
With a cutoff value of 0.05, we are left with 11 attributes:

Table 2

Attribute Name	Pearson Correlation Coefficient
Driver Distracted By	0.3668
Vehicle First Impact Location	0.20997
Vehicle Damage Extent	0.12196
Speed Limit	0.11367
Parked Vehicle	0.10947
Injury Severity	0.10293
Vehicle Movement	0.10077
Traffic Control	0.07267
Collision Type	0.06116
Vehicle Year	0.05368
ACRS Report Type	0.05065

CfsSubsetEval

CfsSubsetEval is an attribute selection algorithm that prioritizes selecting attributes that have a high predictive ability while also reducing redundancy between the attributes. It selects attributes that have a high correlation with the class and low intercorrelation with other attributes.

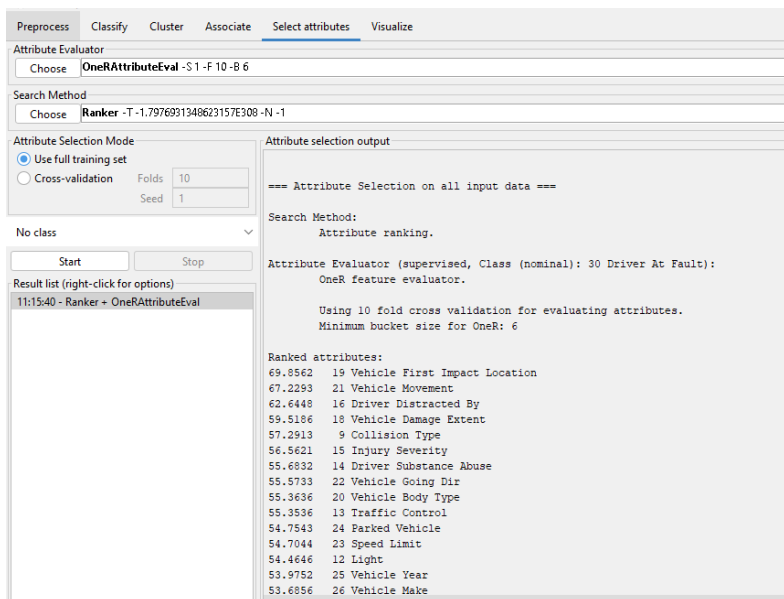


Based the results of CfsSubset attribute selection algorithm, we selected three attributes:

- Report Number
- Driver Distracted By
- Vehicle Movement

OneRAttributeEval

OneRAttributeEval is an attribute selection algorithm that evaluates attributes by using the OneR classifier. It creates a ruleset based on each attribute and compares the error rates between each ruleset. OneRAttributeEval ranks attributes that create a ruleset with a low error rate and a high accuracy above attributes that create a ruleset with a high error rate and a low accuracy.



With a cutoff of 54, we are left with 13 attributes:

Table 3

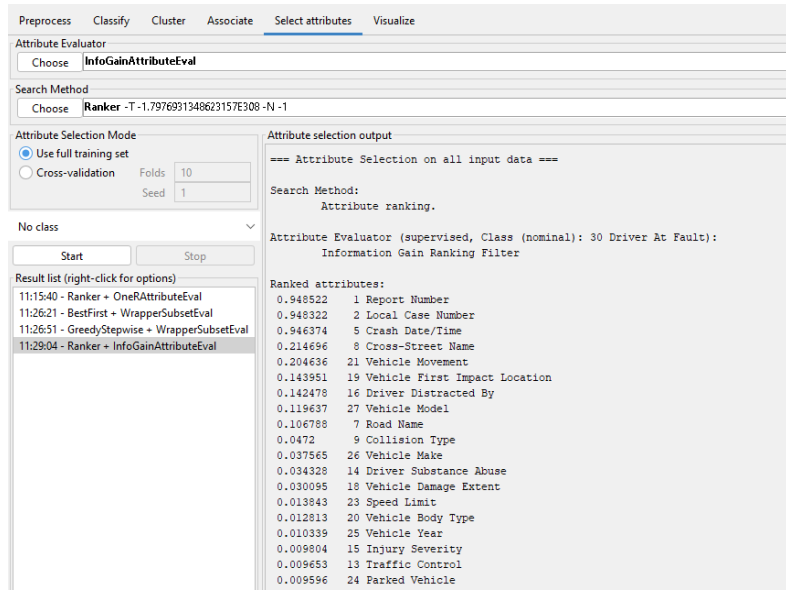
Attribute Name	OneR Value
Vehicle First Impact Location	69.8562
Vehicle Movement	67.2293
Driver Distracted By	62.6448
Vehicle Damage Extent	59.5186
Collision Type	57.2913
Injury Severity	56.5621
Driver Substance Abuse	55.6832
Vehicle Going Dir	55.5733
Vehicle Body Type	55.3636
Traffic Control	55.3536
Parked Vehicle	54.7543
Speed Limit	54.7044
Light	54.4646

InfoGainAttributeEval

InfoGainAttributeEval is an attribute selection algorithm that selects attributes based on how much information is gained from that attribute with respect to the class. The information gained from an attribute is defined as:

$$\text{InfoGain}(\text{Class}, \text{Attribute}) = \text{Entropy}(\text{Class}) - \text{Entropy}(\text{Class}|\text{Attribute})$$

This means that the information gained is equal to reduction in uncertainty of the class as a result of knowing the class attribute. InfoGainAttributeEval ranks attributes with a greater InfoGain value higher than attributes with a lower InfoGain value.



With a cutoff value of 0.1, we are left with 9 attributes:

Table 4

Attribute Name	Info Gain
Report Number	0.948522
Local Case Number	0.948322
Crash Date/Time	0.946374
Cross-Street Name	0.214696
Vehicle Movement	0.204636
Vehicle First Impact Location	0.143951
Driver Distracted By	0.142478
Vehicle Model	0.119637
Road Name	0.106788

Intuition Attribute Selection

We chose to keep the following attributes because we believe that each one has some correlation to the class, Driver at Fault.

- ACRS Report Type
- Route Type
- Road Name
- Cross-Street Name
- Collision Type
- Weather
- Surface Condition
- Light
- Traffic Control
- Driver Substance Abuse
- Injury Severity
- Driver Distracted By
- Vehicle Damage Extent
- Vehicle First Impact Location
- Vehicle Body Type
- Vehicle Movement
- Speed Limit
- Parked Vehicle

Results

We evaluated the performance of the following classifiers on our dataset:

- NaïveBayes
- KStar
- DecisionTable
- OneR

NaïveBayes

The NaïveBayes classifier works by assuming that all attributes are independent of each other. The calculations for the probability of an attribute corresponding to the class is calculated separately and then combined. To make a prediction, NaïveBayes calculates the probability of each class value by combining the likelihood of the attribute values.

KStar

KStar classifier makes predictions by memorizing training instances and comparing how similar a new testing instance is to previously existing training instances. Then, it makes a prediction for the testing instance based on the class value of its closest “neighbor.”

DecisionTable

DecisionTable is used for building and using a simple decision table majority classifier. It is a compact way of recording the different actions made in different sets of conditions to reach decisions. The classifier finds the combinations of attribute values that occur the most times to create rule sets. It then uses these rules to predict class values for the training instances. If a combination of attribute values was not previously noted, the prediction is then the mode of the class value.

OneR

OneR classifier works by creating one rule that determines the predictions of the model. The classifier creates a rule for each attribute that corresponds to the rule with the smallest error rate. Then, the OneR classifier compares the error rates for all rules of all attributes and chooses the one with the least error rate to determine the predictions for the entire model.

CorrelationAttributeEval with NaïveBayes Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.36 seconds

=== Summary ===

Correctly Classified Instances      765      76.3473 %
Incorrectly Classified Instances    237      23.6527 %
Kappa statistic                    0.5302
Mean absolute error                 0.2699
Root mean squared error             0.3859
Relative absolute error             54.2456 %
Root relative squared error         77.3766 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.837   0.300   0.708     0.837   0.767     0.538   0.876    0.879    NO
      0.700   0.163   0.831     0.700   0.760     0.538   0.876    0.879    YES
Weighted Avg.  0.763   0.227   0.774     0.763   0.763     0.538   0.876    0.879

=== Confusion Matrix ===

  a  b  <-- classified as
390 76 |  a = NO
161 375 |  b = YES

```

CfsSubsetEval with NaïveBayes Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.3 seconds

=== Summary ===

Correctly Classified Instances      743      74.1517 %
Incorrectly Classified Instances    259      25.8483 %
Kappa statistic                    0.4914
Mean absolute error                0.3238
Root mean squared error            0.4016
Relative absolute error             65.0764 %
Root relative squared error         80.5252 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.886   0.384   0.667     0.886   0.761     0.515   0.823    0.778    NO
          0.616   0.114   0.862     0.616   0.718     0.515   0.823    0.832    YES
Weighted Avg.   0.742   0.240   0.771     0.742   0.738     0.515   0.823    0.807

=== Confusion Matrix ===

  a    b  <-- classified as
413  53 |  a = NO
206 330 |  b = YES

```

OneRAttributeEval with NaïveBayes Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.44 seconds

=== Summary ===

Correctly Classified Instances      788      78.6427 %
Incorrectly Classified Instances    214      21.3573 %
Kappa statistic                    0.5655
Mean absolute error                0.2546
Root mean squared error            0.3769
Relative absolute error             51.1787 %
Root relative squared error         75.5634 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.678   0.119   0.832     0.678   0.747     0.574   0.881    0.879    NO
          0.881   0.322   0.759     0.881   0.815     0.574   0.881    0.893    YES
Weighted Avg.   0.786   0.228   0.793     0.786   0.784     0.574   0.881    0.886

=== Confusion Matrix ===

  a    b  <-- classified as
316 150 |  a = NO
 64 472 |  b = YES

```

InfoGainAttributeEval with NaïveBayes Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.16 seconds

=== Summary ===

Correctly Classified Instances      737          73.5529 %
Incorrectly Classified Instances    265          26.4471 %
Kappa statistic                    0.4679
Mean absolute error                 0.2999
Root mean squared error             0.4093
Relative absolute error             60.2732 %
Root relative squared error         82.0536 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
      0.708    0.241    0.719    0.708    0.714    0.468  0.836    0.841    NO
      0.759    0.292    0.750    0.759    0.754    0.468  0.836    0.841    YES
Weighted Avg.   0.736    0.268    0.735    0.736    0.735    0.468  0.836    0.841

=== Confusion Matrix ===

  a  b  <-- classified as
330 136 | a = NO
129 407 | b = YES

```

Intuition Attribute Selection with NaïveBayes Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.28 seconds

=== Summary ===

Correctly Classified Instances      777          77.5449 %
Incorrectly Classified Instances    225          22.4551 %
Kappa statistic                    0.5469
Mean absolute error                 0.2608
Root mean squared error             0.3905
Relative absolute error             52.4246 %
Root relative squared error         78.3011 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
      0.727    0.183    0.776    0.727    0.751    0.548  0.867    0.871    NO
      0.817    0.273    0.775    0.817    0.796    0.548  0.867    0.874    YES
Weighted Avg.   0.775    0.231    0.775    0.775    0.775    0.548  0.867    0.873

=== Confusion Matrix ===

  a  b  <-- classified as
339 127 | a = NO
 98 438 | b = YES

```

CorrelationAttributeEval with KStar Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 10.66 seconds

=== Summary ===

Correctly Classified Instances      810          80.8383 %
Incorrectly Classified Instances    192          19.1617 %
Kappa statistic                    0.6145
Mean absolute error                 0.28
Root mean squared error             0.3711
Relative absolute error             56.2834 %
Root relative squared error         74.3961 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.785   0.172   0.799     0.785   0.792     0.615   0.883    0.879    NO
          0.828   0.215   0.816     0.828   0.822     0.615   0.883    0.884    YES
Weighted Avg.   0.808   0.195   0.808     0.808   0.808     0.615   0.883    0.882

=== Confusion Matrix ===

  a  b  <-- classified as
366 100 |  a = NO
 92 444 |  b = YES

```

CfsSubsetEval with KStar Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 2.76 seconds

=== Summary ===

Correctly Classified Instances      743          74.1517 %
Incorrectly Classified Instances    259          25.8483 %
Kappa statistic                    0.4925
Mean absolute error                 0.4129
Root mean squared error             0.4329
Relative absolute error             82.9893 %
Root relative squared error         86.791 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.903   0.399   0.663     0.903   0.765     0.522   0.824    0.780    NO
          0.601   0.097   0.877     0.601   0.713     0.522   0.824    0.831    YES
Weighted Avg.   0.742   0.237   0.778     0.742   0.737     0.522   0.824    0.807

=== Confusion Matrix ===

  a  b  <-- classified as
421  45 |  a = NO
214 322 |  b = YES

```


OneRAttributeEval with KStar Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.12 seconds

=== Summary ===

Correctly Classified Instances      697          69.5609 %
Incorrectly Classified Instances    305          30.4391 %
Kappa statistic                    0.3729
Mean absolute error                 0.3044
Root mean squared error            0.5517
Relative absolute error             61.176 %
Root relative squared error        110.6136 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.485   0.121   0.777    0.485   0.597     0.400   0.682    0.616    NO
          0.879   0.515   0.662    0.879   0.755     0.400   0.682    0.647    YES
Weighted Avg.   0.696   0.332   0.716    0.696   0.682     0.400   0.682    0.633

=== Confusion Matrix ===

  a  b  <-- classified as
226 240 |  a = NO
 65 471 |  b = YES

```

InfoGainAttributeEval with KStar Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 20.74 seconds

=== Summary ===

Correctly Classified Instances      745          74.3513 %
Incorrectly Classified Instances    257          25.6487 %
Kappa statistic                    0.4839
Mean absolute error                 0.3609
Root mean squared error            0.4326
Relative absolute error             72.5343 %
Root relative squared error        86.7282 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.715   0.231   0.729    0.715   0.722     0.484   0.797    0.724    NO
          0.769   0.285   0.756    0.769   0.762     0.484   0.797    0.760    YES
Weighted Avg.   0.744   0.260   0.743    0.744   0.743     0.484   0.797    0.743

=== Confusion Matrix ===

  a  b  <-- classified as
333 133 |  a = NO
124 412 |  b = YES

```

Intuition Attribute Selection with KStar Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 45.66 seconds

=== Summary ===

Correctly Classified Instances      753          75.1497 %
Incorrectly Classified Instances    249          24.8503 %
Kappa statistic                    0.502
Mean absolute error                 0.3022
Root mean squared error             0.4177
Relative absolute error             60.7348 %
Root relative squared error         83.7349 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
              0.755   0.252   0.723     0.755   0.739     0.502   0.824    0.789    NO
              0.748   0.245   0.779     0.748   0.763     0.502   0.824    0.812    YES
Weighted Avg.  0.751   0.248   0.753     0.751   0.752     0.502   0.824    0.801

=== Confusion Matrix ===
  a  b  <-- classified as
352 114 |  a = NO
135 401 |  b = YES

```

CorrelationAttributeEval with DecisionTable Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.27 seconds

=== Summary ===

Correctly Classified Instances      806          80.4391 %
Incorrectly Classified Instances    196          19.5609 %
Kappa statistic                    0.6059
Mean absolute error                 0.286
Root mean squared error             0.3703
Relative absolute error             57.4794 %
Root relative squared error         74.249 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
              0.770   0.166   0.801     0.770   0.786     0.606   0.886    0.884    NO
              0.834   0.230   0.807     0.834   0.820     0.606   0.886    0.881    YES
Weighted Avg.  0.804   0.200   0.804     0.804   0.804     0.606   0.886    0.882

=== Confusion Matrix ===
  a  b  <-- classified as
359 107 |  a = NO
 89 447 |  b = YES

```

CfsSubsetEval with with DecisionTable Classifier

```

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances      747          74.5509 %
Incorrectly Classified Instances    255          25.4491 %
Kappa statistic                    0.4998
Mean absolute error                 0.3315
Root mean squared error             0.3998
Relative absolute error             66.6296 %
Root relative squared error        80.1572 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.899   0.388   0.668     0.899   0.767     0.527   0.829    0.781    NO
          0.612   0.101   0.875     0.612   0.720     0.527   0.829    0.833    YES
Weighted Avg.   0.746   0.234   0.779     0.746   0.742     0.527   0.829    0.809

=== Confusion Matrix ===

  a  b  <-- classified as
419 47 |  a = NO
208 328 |  b = YES

```

OneRAttributeEval with DecisionTable Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 2.89 seconds

=== Summary ===

Correctly Classified Instances      803          80.1397 %
Incorrectly Classified Instances    199          19.8603 %
Kappa statistic                    0.6016
Mean absolute error                 0.2886
Root mean squared error             0.3719
Relative absolute error             57.9935 %
Root relative squared error        74.562 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.800   0.198   0.779     0.800   0.789     0.602   0.886    0.883    NO
          0.802   0.200   0.822     0.802   0.812     0.602   0.886    0.882    YES
Weighted Avg.   0.801   0.199   0.802     0.801   0.802     0.602   0.886    0.882

=== Confusion Matrix ===

  a  b  <-- classified as
373 93 |  a = NO
106 430 |  b = YES

```

InfoGainAttributeEval with DecisionTable Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.37 seconds

=== Summary ===

Correctly Classified Instances      797          79.5409 %
Incorrectly Classified Instances    205          20.4591 %
Kappa statistic                    0.5853
Mean absolute error                 0.2912
Root mean squared error             0.3694
Relative absolute error             58.5313 %
Root relative squared error         74.0592 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.715   0.134   0.822     0.715   0.765     0.590   0.881    0.876    NO
          0.866   0.285   0.777     0.866   0.819     0.590   0.881    0.880    YES
Weighted Avg.   0.795   0.215   0.798     0.795   0.794     0.590   0.881    0.878

=== Confusion Matrix ===

  a  b  <-- classified as
333 133 |  a = NO
 72 464 |  b = YES

```

Intuition Attribute Selection with DecisionTable Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.47 seconds

=== Summary ===

Correctly Classified Instances      803          80.1397 %
Incorrectly Classified Instances    199          19.8603 %
Kappa statistic                    0.6016
Mean absolute error                 0.2886
Root mean squared error             0.3719
Relative absolute error             57.9935 %
Root relative squared error         74.562 %
Total Number of Instances          1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.800   0.198   0.779     0.800   0.789     0.602   0.886    0.883    NO
          0.802   0.200   0.822     0.802   0.812     0.602   0.886    0.882    YES
Weighted Avg.   0.801   0.199   0.802     0.801   0.802     0.602   0.886    0.882

=== Confusion Matrix ===

  a  b  <-- classified as
373  93 |  a = NO
106 430 |  b = YES

```

CorrelationAttributeEval with OneR Classifier

```

=== Evaluation on test set ===

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

=== Summary ===

Correctly Classified Instances      697          69.5609 %
Incorrectly Classified Instances    305          30.4391 %
Kappa statistic                    0.3729
Mean absolute error                 0.3044
Root mean squared error             0.5517
Relative absolute error             61.176 %
Root relative squared error        110.6136 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.485    0.121    0.777     0.485    0.597      0.400    0.682    0.616     NO
          0.879    0.515    0.662     0.879    0.755      0.400    0.682    0.647     YES
Weighted Avg.    0.696    0.332    0.716     0.696    0.682      0.400    0.682    0.633

=== Confusion Matrix ===

  a    b  <-- classified as
226 240 |  a = NO
 65 471 |  b = YES

```

CfsSubsetEval with OneR Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.22 seconds

=== Summary ===

Correctly Classified Instances      451          45.01 %
Incorrectly Classified Instances    551          54.99 %
Kappa statistic                    -0.03
Mean absolute error                 0.5499
Root mean squared error             0.7416
Relative absolute error            110.5179 %
Root relative squared error        148.6738 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.968    1.000    0.457     0.968    0.621     -0.132    0.484    0.457     NO
          0.000    0.032    0.000     0.000    0.000     -0.132    0.484    0.535     YES
Weighted Avg.    0.450    0.482    0.213     0.450    0.289     -0.132    0.484    0.499

=== Confusion Matrix ===

  a    b  <-- classified as
 451  15 |  a = NO
 536   0 |  b = YES

```

OneRAttributeEval with OneR Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.11 seconds

=== Summary ===

Correctly Classified Instances      697          69.5609 %
Incorrectly Classified Instances    305          30.4391 %
Kappa statistic                    0.3729
Mean absolute error                 0.3044
Root mean squared error             0.5517
Relative absolute error             61.176 %
Root relative squared error        110.6136 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
0.485    0.121    0.777    0.485    0.597    0.400    0.682    0.616    NO
0.879    0.515    0.662    0.879    0.755    0.400    0.682    0.647    YES
Weighted Avg.    0.696    0.332    0.716    0.696    0.682    0.400    0.682    0.633

=== Confusion Matrix ===

  a    b  <-- classified as
226 240 |  a = NO
 65 471 |  b = YES

```

InfoGainAttributeEval with OneR Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.28 seconds

=== Summary ===

Correctly Classified Instances      451          45.01 %
Incorrectly Classified Instances    551          54.99 %
Kappa statistic                    -0.03
Mean absolute error                 0.5499
Root mean squared error             0.7416
Relative absolute error            110.5179 %
Root relative squared error        148.6738 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

              TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
0.968    1.000    0.457    0.968    0.621    -0.132    0.484    0.457    NO
0.000    0.032    0.000    0.000    0.000    -0.132    0.484    0.535    YES
Weighted Avg.    0.450    0.482    0.213    0.450    0.289    -0.132    0.484    0.499

=== Confusion Matrix ===

  a    b  <-- classified as
451  15 |  a = NO
536   0 |  b = YES

```

Intuition Attribute Selection with OneR Classifier

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 1.14 seconds

=== Summary ===

Correctly Classified Instances      697          69.5609 %
Incorrectly Classified Instances    305          30.4391 %
Kappa statistic                    0.3729
Mean absolute error                 0.3044
Root mean squared error             0.5517
Relative absolute error             61.176 %
Root relative squared error        110.6136 %
Total Number of Instances         1002

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
0.485    0.121    0.777    0.485    0.597    0.400    0.682    0.616    NO
0.879    0.515    0.662    0.879    0.755    0.400    0.682    0.647    YES
Weighted Avg.    0.696    0.332    0.716    0.696    0.682    0.400    0.682    0.633

=== Confusion Matrix ===
  a  b  <-- classified as
226 240 |  a = NO
 65 471 |  b = YES

```

Table 5

Accuracy (%)		Attribute Selection Algorithms				
		Correlation	CfsSubset	OneR	InfoGain	Intuition
Classifiers	NaïveBayes	76.3473	74.1517	78.6427	73.5529	77.5449
	KStar	80.8383	74.1517	69.5609	74.3513	75.1497
	DecisionTable	80.4391	74.5509	80.1397	79.5409	80.1397
	OneR	69.5609	45.01	69.5609	45.01	69.5609

Table 6

True Positive Rate		Attribute Selection Algorithms				
		Correlation	CfsSubset	OneR	InfoGain	Intuition
Classifiers	NaïveBayes	0.700	0.616	0.881	0.759	0.817
	KStar	0.828	0.601	0.879	0.769	0.748
	DecisionTable	0.834	0.612	0.802	0.866	0.802
	OneR	0.879	0.000	0.879	0.000	0.879

Table 7

False Positive Rate		Attribute Selection Algorithms				
		Correlation	CfsSubset	OneR	InfoGain	Intuition
Classifiers	NaïveBayes	0.163	0.114	0.322	0.292	0.273
	KStar	0.215	0.097	0.515	0.285	0.245
	DecisionTable	0.230	0.101	0.200	0.285	0.200
	OneR	0.515	0.032	0.515	0.032	0.515

Table 8

ROC Area		Attribute Selection Algorithms				
		Correlation	CfsSubset	OneR	InfoGain	Intuition
Classifiers	NaïveBayes	0.876	0.823	0.881	0.836	0.867
	KStar	0.883	0.824	0.682	0.797	0.824
	DecisionTable	0.886	0.829	0.882	0.881	0.886
	OneR	0.682	0.484	0.682	0.484	0.682

The OneR classifier performed poorly in comparison to the other classifier algorithms, especially when paired with the CfsSubset attribute selection algorithm or InfoGain attribute selection algorithm. We believe that this occurred because the OneR classifier chooses one attribute as the rule to predict the class, which may be too simple to accurately predict the class. The reason that OneR performs especially poorly when paired with CfsSubset or InfoGain may be because the attributes selected by these attribute selection algorithms have low correlation with other attributes to reduce redundancy, but also appear to have low correlation with the class variable. Thus, the OneR classifier was unable to choose an attribute that correlated strongly with the class to create its ruleset.

Conclusion

We think that the DecisionTable model with CorrelationAttributeEval is best for our data mining goals. This is because it had consistently good results among the performance measures we inspected: accuracy, true positive rate, false positive rate, and ROC Curve. This model had an accuracy of 80.44%, the highest accuracy of all models. It also had good results for the other performance measures. This model had among the highest true positive rates and one of the highest ROC Curve areas.

This model could be improved by optimizing the cutoff value used for selecting attributes after performing CorrelationAttributeEval. We arbitrarily selected a cutoff value of 0.05; however, if an algorithm was used to find the optimal cutoff value, the attributes selected could be more effective at predicting the class. Through this project, we gained experience in using the Weka application with real world data that holds imperfections, such as formatting issues and missing values. Though we previously learned in class the steps to take to make a machine learning model in Weka, actually doing so with an imperfect dataset was enlightening. We had to figure out how to solve little issues that would cause programs to not run at all. Additionally, we learned about the potential flaws of some attribute selection algorithms, including CfsSubset. This attribute selection algorithm attempts to reduce redundancy by selecting attributes that have low intercorrelation; however, we noticed that while this algorithm reduced redundancy, it also selected attributes that did not have much correlation with the class variable. With this new understanding of the strengths and weaknesses of different attribute selection algorithms, we will be better informed to use an attribute selection algorithm that fits with our data mining goals in the future.

Team Members & Tasks Performed

Finding the Data: Yuan Yin

Building Proposal: Yuan Yin & Julia Chen

Defining Attributes: Yuan Yin & Julia Chen

Preprocessing:

- Remove Instances Missing the Class: Yuan Yin
- Remove Attributes Missing More Than 70% of Their Values: Yuan Yin
- Remove Derived Attributes: Yuan Yin
- Remove Unnecessary Attributes: Yuan Yin
- Stratified Random Sampling: Julia Chen
- Unify Values: Julia Chen
- Fill in Missing Values: Julia Chen

Attribute Selection:

- CorrelationAttributeEval: Yuan Yin

- CfsSubsetEval: Yuan Yin
- InfoGainAttributeEval: Julia Chen
- OneRAttributeEval: Julia Chen
- Intuition Attribute Selection: Julia Chen

Classifiers: Julia Chen & Yuan Yin

Results Analysis: Julia Chen & Yuan Yin

Building Final Report: Julia Chen & Yuan Yin

Steps to Reproduce Our Model: DecisionTable classifier with CorrelationAttributeEval:

1. Open Weka and open the `post_pre_processing_ready_for_attribute_selection.arff` file.
2. Go to the “Select attributes” tab and on the left panel click the button reading “No class” and choose “(Nom) Driver at Fault” as the class from the dropdown menu.
3. Select “CorrelationAttributeEval” as Attribute Evaluator and “Ranker” as the Search Method and hit “Start.”
4. Using 0.05 as the cutoff value, take note of all attributes with a Pearson Correlation Coefficient greater than the cutoff value.
5. Open the `train.arff` file in Weka and remove all attributes except the ones recorded in Step 4 and the class (Driver at Fault).
6. Save this file as `train_correlation.arff`.
7. Repeat step 5 for the `test.arff` and save this file as `test_correlation.arff`.
8. Open Weka Explorer and load the `train_correlation.arff` in the “Preprocess” tab.
9. Select on the “Classify” tab and click “Supplied test set” in the left panel under “Test options”.
10. Click to open the `test_correlation.arff`.
11. Select the class (Driver at Fault).
12. Click “Choose” on the “Classifier” selection and select the “DecisionTable” classifier (weka → classifiers → rules → DecisionTable).
13. Click “Start”.