Machine Learning - Dr. Yilmaz Period 5

Predicting Crash Type from Crash Report Incident

Team Members: Gabriel Xu, Andrew Chen

10/07/2024

Table of Contents

Part 1: Dataset Overview

Section 1.1: Introduction

The dataset we used is the “Crash Data” dataset, linked here: https://catalog.data.gov/dataset/crash-data. There are 24241 instances of 42 attributes of crashes occurring in the Town of Cary, North Carolina. We plan to classify **how severe** the results of the crash were. This means predicting whether there were no injuries or fatalities, injuries but no fatalities, or fatalities. This dataset will be useful for finding how certain **conditions** such as location, weather, and the road combine to **impact** how **dangerous** a crash could be.

Section 1.2: Meaning of Attributes

1. tamainid

- **Meaning**: Unique ID for each traffic accident entry.
- **Example**: 48247, 48253 – these are unique IDs for individual accidents

2. location_description

- **Meaning**: Describes the location of the accident (e.g., street names or distances from landmarks).
- **Example**: “30 FEET FROM SR3977 (SW CARY PKWY)” – this accident occurred 30 feet from a specific street,

3. rdfeature

- **Meaning**: Describes special road features near the accident location.
- **Example**: “RAILROAD CROSSING” – The accident took place near a railroad crossing.

4. rdcharacter

- **Meaning**: The road’s physical characteristics, such as whether it’s straight or curved.
- **Example**: “STRAIGHT, LEVEL” – the road was straight and level at the time of the accident.

5. rdclass

- **Meaning**: Classification of the road, such as whether it’s a local street or highway.
- **Example**: “LOCAL STREET” – the accident occurred on a smaller, local street

6. rdconfigur

- **Meaning**: Describes the configuration of the road, such as whether it’s divided or undivided.
- **Example**: “TWO-WAY, DIVIDED, POSITIVE MEDIAN BARRIER” – the road has two-way traffic with a median barrier.

7. rdsurface

- **Meaning**: The type of surface the road has.
- **Example**: “SMOOTH ASPHALT” – the road surface was smooth asphalt

8. rdcondition

- **Meaning**: The condition of the road at the time of the accident.
- **Example**: “DRY” – the road was dry when the accident occurred.

9. lightcond

- **Meaning**: Describes the lighting conditions at the time of the accident.
- **Example**: “DAYLIGHT” – the accident occurred during daylight hours

10. weather

- **Meaning**: Weather conditions during the time of the accident.
- **Example**: “CLEAR” – the weather was clear during the accident

11. trafcontrl

- **Meaning**: Describes the traffic control present at the location (e.g., traffic lights, stop signs).
- **Example**: “RR GATE AND FLASHER” – A railroad crossing gate and flasher were present.

12. lat

- **Meaning**: Latitude coordinate of the accident location.
- **Example**: -78.821706 – the latitude of the accident site

13. lon

- **Meaning**: Longitude coordinate of the accident location.
- **Example**: 35.761999 – the longitude of the accident site

14. lon2

- **Meaning**: A second longitude value, possibly for marking a different point in the accident area.
- **Example**: -78.787907 – Another longitude point related to the accident.

15. lat2

- **Meaning**: A second latitude value, similar to lon2 for another geographic point in the accident.
- **Example**: 35.716448 – Another latitude point.

16. tract

- **Meaning**: A geographic subdivision, probably a census tract
- **Example**: “P054” – A specific census tract for geographic reference.

17. zone

- **Meaning**: Refers to a traffic or urban zone (e.g., residential, school zone).
- **Example**: “116” – Numeric representation of the zone where the accident occurred.

18. fatality

- **Meaning**: Indicates if there was a fatality (Yes/No).
- **Example**: “No” – No fatalities were recorded for this accident.

19. possblinj

- **Meaning**: Indicates if there were possible injuries (Yes/No).
- **Example**: “No” – No injuries were recorded for this accident.

20. numpassengers

- **Meaning**: The number of passengers in the vehicles involved in the accident
- **Example**: “1” – There was one passenger involved.

21. numpedestrians

- **Meaning**: The number of pedestrians involved in the accident.
- **Example**: “0” – No pedestrians were involved in the accident.

22. contrcir1_desc

- **Meaning**: Describes a contributing factor to the accident (e.g., distracted driving, speeding).
- **Example**: “NONE” – No specific contributing factor was recorded.

23. contrcir2_desc
- **Meaning**: Describes an additional contributing factor, if applicable.
- **Example**: “NONE” – No secondary contributing factor was recorded.

24. contrcir3_desc
- **Meaning**: Describes a third contributing factor, if applicable.
- **Example**: “NONE” – No third contributing factor was recorded.

25. contrcir4_desc
- **Meaning**: Describes a fourth contributing factor, if applicable.
- **Example**: “NONE” – No fourth contributing factor was recorded.

26. vehicle1
- **Meaning**: Describes the first vehicle involved in the accident.
- **Example**: “PICKUP” – The first vehicle was a pickup truck.

27. vehicle2
- **Meaning**: Describes the second vehicle involved, if applicable.
- **Example**: “SPORT UTILITY” – The second vehicle was a sport utility vehicle.

28. vehicle3
- **Meaning**: Describes the third vehicle involved, if applicable.
- **Example**: “PASSENGER CAR” – The third vehicle was a passenger car.

29. vehicle4
- **Meaning**: Describes the fourth vehicle involved, if applicable.
- **Example**: “None” – No fourth vehicle was involved.

30. vehicle5
- **Meaning**: Describes the fifth vehicle involved, if applicable.
- **Example**: “None” – No fifth vehicle was involved.

31. workarea
- **Meaning**: Indicates whether the accident occurred in a work zone (Yes/No).
- **Example**: “NO” – The accident did not occur in a work zone.

32. records
- **Meaning**: Could refer to the record number or entry in the database.
- **Example**: “10003” – This is the database record number.

33. ta_date
- **Meaning**: The date when the traffic accident occurred.
- **Example**: “2021-07-07” – The accident occurred on July 7, 2021.

34. ta_time
- **Meaning**: The time when the traffic accident occurred.
- **Example**: “2:18:32 PM” – The accident occurred at this specific time.

35. crash_date
- **Meaning**: The timestamp for when the crash was officially recorded.
- **Example**: “2021-07-07T18:18:32+00:00” – The crash was recorded at this time in UTC format.

36. geo_location

- **Meaning**: The latitude and longitude of the crash combined for easy geographic reference.
- **Example**: “35.716440073, -78.78796424” – The location of the accident.

37. year
- **Meaning**: The year in which the accident occurred.
- **Example**: “2021” – The accident occurred in 2021.

38. fatalities
- **Meaning**: If there were fatalities in the accident.
- **Example**: “No” – No fatalities resulted from this accident.

39. injuries
- **Meaning**: If there were injuries in the accident.
- **Example**: “No” – No injuries were reported.

40. month
- **Meaning**: The month when the accident occurred.
- **Example**: “7” – The accident occurred in July.

41. contributing_factor
- **Meaning**: Main contributing factor that caused the accident.
- **Example**: “NONE, NONE” – No contributing factors were recorded.

42. vehicle_type
- **Meaning**: Types of vehicles involved in the accident.
- **Example**: “PICKUP” – A pickup truck was involved in the accident.

Section 1.3: Preprocessing Plans

Most of the columns with categorical data types are all skewed towards certain values. The class variables, the number of fatalities and injuries are also heavily skewed towards lower values. For preprocessing we will need to fix areas such as missing values and deleting columns that won’t be helpful. The columns and rows with a high amount of missing values (greater than ~70%) will be deleted and the remaining ones will likely have missing values replaced. Another thing to fix would be how certain features contain a list of individual characteristics that we should further separate. Another obvious step is to normalize the data since the attributes are on different scales.

Part 2: Preprocessing

Section 2.1: Transform Class Column

Using the columns of “fatalities” and “injuries”, construct a new “**class**” column with possible values of “crash”, “injury”, or “fatalities”. If “fatalities” is Yes, the value will be “fatalities”. Then if “injuries” is Yes, the value will be “injury”. Otherwise, the value will be “crash”. We then obviously removed the “injuries” and “fatalities” features. This dataset is titled as “cpd-crash-incidents.csv” in the Google Drive folder.

Section 2.2: Delete Useless Columns and Rows

We dropped the columns “tamainid” and “records” because they likely don't have any correlation with the class and are just IDs. The “fatality” and “posiblinj” columns were removed because they are redundancies of the “fatalities” and “injuries” columns. “lat” and “lon” were likewise removed because they are repeats of “lat2” and “lon2”, and they have much more missing values than their counterparts. “location_description”, “tract”, “contributing_factor”, and “vehicle_type” were removed because there aren't a lot of repeats among instances as there are a lot of different possible values, so basic classifier models wouldn’t be able to properly process this text. “ta_date”, “crash_date”, and “year” for similar reasons as above. “geo_location” is removed since it is derived from longitude and latitude. Although “zone” is technically derived from the location to some extent, we thought it was worth keeping for now since the relationship is not a clear, direct one.

We also checked for repeat instances and removed them.

The code to complete Sections 2.1 and 2.2 is here:

```
data=[row.split(';') for row in
open("cpd-crash-incidents.csv").read().strip().split("\n")]
new_data=[]
to_delete=["tamainid", "records", "fatality", "year",
"possblinj", "lat", "lon", "location_description", "tract",
"contributing_factor", "vehicle_type", "ta_date", "crash_date",
"geo_location"]
for row in data:
  a=[]
  for i, x in enumerate(row):
    if data[0][i] not in to_delete:
      a.append(x)
  if a[-3]=="Yes": a.append("fatalities")
  elif a[-2]=="Yes": a.append("injury")
  else: a.append("crash")
  a=a[:-4]+a[-2:]
```

```
  new_data.append(a)
new_data[0][-1]="class"
```

Section 2.3: Clean up N/A/Missing Values

All the cells with a value of N/A, “None”, “Other”, and “Unknown”, were converted to an empty cell. We decided to substitute missing values in the “numpedestrians” column with the value of 0.

The code to do this is here:

```
for i in range(len(new_data)):
  for j in range(len(new_data[0])):
    if new_data[i][j] in ["NONE", "OTHER *", "UNKNOWN"]:
      new_data[i][j]=""
    if new_data[i][j]=="" and j==13:
      new_data[i][j]="0"
```

Section 2.4: Transform Attributes into Usable Form

We added a new column “vehicles” based on the number of values for “vehicle1” through “vehicle5” are not empty.

```
for i in range(1, len(new_data)):
  count=0
  for j in range(len(new_data[0])):
    if new_data[0][j][:-1]=="vehicle" and new_data[i][j]!="":
      count+=1
  new_data[i]=new_data[i][:-1]+[count]+new_data[i][-1:]
new_data[0]=new_data[0][:-1]+["vehicles"]+new_data[0][-1:]
```

The “ta_time” column was transformed into groups for each hour of the day, as minutes are too specific for models to use.

```
for i in range(1, len(new_data)):
  for j in range(len(new_data[0])):
    if new_data[0][j]=="ta_time":
      new_time=int(new_data[i][j][:new_data[i][j].index(':')])
      if new_data[i][j][-2:]=="PM":
        if new_time!=12:
          new_time+=12
      elif new_time==12:
        new_time=0
      new_data[i][j]=new_time
with open("final.csv", "w") as f:
```

```
for row in new_data:
  f.write(";".join(map(str, row))+"\n")
```

Section 2.5: Handle Empty Cells

The columns and rows with more than >=70% missing values were dropped. First, we removed the attributes of “contrcir1_desc”, “contrcir2_desc”, “contrcir3_desc”, “contrcir4_desc”, “vehicle3”, “vehicle4”, and “vehicle5”.

There were no instances we needed to delete.

Then we ran WEKA’s method filter of replacing missing values with the mode/mean of that attribute.

Section 2.6: Normalize Data

We then decided to use z-score normalization on all the discrete attributes through the standardize filter (except for the class).

Section 2.8: Summary

After all of these preprocessing steps, the dataset (the one titled “final.csv”) now has 24241 instances of 20 columns (not including class). For the class labels, 0.18% are “fatalities”, 13.34% are “injury”, and 86.47% are “crash”.

Part 3: Attribute Analysis

A. non-Weka Selection Method

After careful consideration, we chose to keep the following attributes, as they would likely have a considerable impact on crash severity:

1) **weather** – weather conditions (e.g. rainy or snowy) can impact visibility and road conditions, leading to more severe accidents
2) **trafcontrl** -- the presence of traffic controls (e.g. stop signs) impacts crash severity
3) **lightcond** -- poor lighting conditions (night or dawn/dusk) could increase crash severity
4) **rdcondition** (Road Condition) – slippery or damaged roads can lead to increased crash severity
5) **rdfeature** – road features like curves, intersections, and narrow lanes may lead to more severe crashes
6) **vehicle1** and **vehicle2** – types of vehicles involved (e.g., cars, trucks, motorcycles) could impact the severity, with larger vehicles likely being more severe
7) **numpassengers** – a higher number of passengers could influence the overall impact as well as severity
8) **numpedestrians** – the presence of pedestrians likely increases the likelihood of severe injury.

B. CorrelationAttributeEval

We used Weka for this approach. This approach evaluates attributes and ranks them based on their individual correlations with the class label (crash, injury, fatalities). We only kept attributes with a **correlation score of 0.02 or higher**. This threshold ensures that the most relevant features are selected while also minimizing noise from less meaningful attributes.

The attributes that remain after applying this selection algorithm are:

1	rdfeature
2	rdcharacter
3	rdclass
4	rdconfigur
5	rdcondition
6	lightcond
7	trafcontrl
8	zone
9	numpedestrians
10	vehicles
11	class

C. InfoGainAttributeEval

We used Weka for this approach. This approach ranks attributes based on how much information they provide about the class label (crash, injury, fatalities). In other words, this method measures the **Information Gain** for each attribute.

Information Gain can be calculated by these three steps.

1) **Entropy of dataset D:**

$$\text{Entropy}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Where p_i is the proportion of instances in class C_i in the dataset D, and m is the number of classes.

2) **Entropy after splitting on attribute A**:

$$\text{Entropy}_A(D) = \sum_{v=1}^{V} \frac{|D_v|}{|D|} \cdot \text{Entropy}(D_v)$$

Where D_v is the subset of D where attribute A has value v, and V is the number of distinct values of A.

3) **Information Gain**:

$$IG(A) = \text{Entropy}(D) - \text{Entropy}_A(D)$$

A higher IG for an attribute indicates that it provides more information about the class labels. The attributes that remain after applying this selection algorithm are below. Attributes were only selected if they had **InfoGain of 0.01 or higher**.

1 rdfeature
2 rdclass
3 rdconfigur
4 trafcontrl
5 numpassengers
6 vehicles
7 class

D. ReliefF

We used Weka for this approach. This technique evaluates the importance of each attribute by comparing how similar instances are within the same class (nearest hits) versus how different they are from different classes (nearest misses). The ReliefF Algorithm then ranks attributes based on how well they separate different classes.

For an attribute A, the weight of the attribute is calculated by:

$$W(A) = W(A) - \frac{1}{m}\sum_{i=1}^{m}[diff(A, \text{nearest\_hit}) - diff(A, \text{nearest\_miss})]$$

Where:

m is the number of sampled instances.

nearest_hit refers to the closest instance of the same class.

nearest_miss refers to the closest instance from a different class.

$diff(A, \text{nearest\_hit})$ is the difference in attribute A between the instance and its nearest hit, and similarly for the nearest miss.

We selected attributes with a weight of **0.02 or higher**. The attributes that remain after applying the algorithm are below:

1	rdfeature
2	rdcharacter
3	rdclass
4	rdconfigur
5	rdsurface
6	lightcond
7	weather
8	trafcontrl
9	lon2
10	zone
11	vehicle1
12	vehicle2
13	ta_time
14	vehicles
15	class

E. CfsSubsetEval

We used Weka for this approach, which selects a subset of features based on how well they predict and how little they overlap each other (both relevant and not redundant).

The attributes that remain after applying this model are:

1	rdfeature
2	rdclass
3	rdconfigur
4	trafcontrl
5	numpassengers
6	numpedestrians
7	vehicles

Part 4: Classifier Models

1. bayes.NaiveBayes

Naive Bayes is a probabilistic classifier based on Baye's theorem. Its called "naive" because it assumes that attributes of a dataset are independent and don't affect each other. The algorithm chooses the class with the highest conditional probability.
Naive Bayes uses Bayes' theorem:

$$P(C_i|X) = \frac{P(X|C_i) \cdot P(C_i)}{P(X)}$$

Where $P(C_i|X)$ is the probability of class C_i given the data X. Because it assumes all features as independent, we simplify $P(X|C_i)$ as:

$$P(X|C_i) = P(x_1|C_i) \cdot P(x_2|C_i) \cdots P(x_n|C_i)$$

Then, the class with the highest $P(C_i|X)$ is predicted

2. tree.J48

J48 is an implementation of the C4.5 decision tree algorithm. It recursively splits the dataset and selects the attribute with the highest information gain. Each internal node represents a decision on an attribute, and the leaf nodes represent the final class prediction.

3. tree.RandomForest

RandomForest is a learning method that builds a collection of decision trees, known as a "forest". Each tree is trained on a random subset of the dataset.
Random Forest selects the class based on majority voting:

$$\hat{C} = \text{mode}(C_1, C_2, \ldots, C_T)$$

Where $(C_1, C_2, \ldots, C_T)$ are the predictions from individual trees.
The final class is the one with the highest number of votes from the trees in the forest.

4. rules.DecisionTable

DecisionTable creates a table of rules based on the attributes. Each rule corresponds to a unique set of conditions on the attributes. If multiple rules apply, the classifier selects the class based on majority rule.
The Decision Table then evaluates each condition:

$$R_i : \text{if } A_1 = x_1 \text{ and } A_2 = x_2 \ \ldots \ \text{then } C_i$$

5. rules.OneR

Unlike DecisionTable which creates multiple rules, the OneR classifier creates one rule for each attribute in the data. The algorithm then selects the rule with the smallest error rate, following the below pseudocode.

For each attribute
    For each value of the attribute
        count the frequency of each class

find the most frequent class

make rule: assign that class to this attribute-value

Compute error rate of the rules (of this attribute)

Choose the rules with the smallest error rate

Non-Weka with Naive Bayes

```
=== Summary ===

Correctly Classified Instances       20876               86.1186 %
Incorrectly Classified Instances      3365               13.8814 %
Kappa statistic                          0.0084
Mean absolute error                      0.1424
Root mean squared error                  0.2786
Relative absolute error                 91.1092 %
Root relative squared error             99.6538 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.994    0.989    0.865      0.994    0.925      0.024    0.653     0.916     crash
                 0.009    0.005    0.230      0.009    0.017      0.021    0.652     0.218     injury
                 0.023    0.001    0.038      0.023    0.029      0.028    0.588     0.009     fatalities
Weighted Avg.    0.861    0.856    0.779      0.861    0.803      0.023    0.652     0.821

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20846    96    20 |     a = crash
  3201    29     5 |     b = injury
    42     1     1 |     c = fatalities
```

Non-Weka with J48

```
=== Summary ===

Correctly Classified Instances       20962               86.4733 %
Incorrectly Classified Instances      3279               13.5267 %
Kappa statistic                          0
Mean absolute error                      0.1563
Root mean squared error                  0.2795
Relative absolute error                 99.9747 %
Root relative squared error            100      %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.865      1.000    0.927      ?        0.500     0.865     crash
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.133     injury
                 0.000    0.000    ?          0.000    ?          ?        0.473     0.002     fatalities
Weighted Avg.    0.865    0.865    ?          0.865    ?          ?        0.500     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20962     0     0 |     a = crash
  3235     0     0 |     b = injury
    44     0     0 |     c = fatalities
```

Non-Weka with RandomForest

```
=== Summary ===

Correctly Classified Instances       20648               85.178  %
Incorrectly Classified Instances      3593               14.822  %
Kappa statistic                          0.0467
Mean absolute error                      0.1478
Root mean squared error                  0.2849
Relative absolute error                 94.5756 %
Root relative squared error            101.9359 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.977    0.945    0.869      0.977    0.919      0.066    0.622     0.900     crash
                 0.054    0.023    0.264      0.054    0.090      0.064    0.621     0.199     injury
                 0.000    0.000    0.000      0.000    0.000      -0.001   0.620     0.005     fatalities
Weighted Avg.    0.852    0.820    0.786      0.852    0.807      0.066    0.622     0.805

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20473   486     3 |     a = crash
  3058   175     2 |     b = injury
    41     3     0 |     c = fatalities
```

Non-Weka with DecisionTable:

```
=== Summary ===

Correctly Classified Instances       20954               86.4403 %
Incorrectly Classified Instances      3287               13.5597 %
Kappa statistic                          0.0002
Mean absolute error                      0.1556
Root mean squared error                  0.2786
Relative absolute error                 99.5436 %
Root relative squared error             99.6517 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.999    0.865      1.000    0.927      0.002    0.555     0.884     crash
                 0.001    0.000    0.167      0.001    0.001      0.002    0.556     0.164     injury
                 0.000    0.000    ?          0.000    ?          ?        0.507     0.004     fatalities
Weighted Avg.    0.864    0.864    ?          0.864    ?          ?        0.555     0.786

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20952    10     0 |     a = crash
  3233     2     0 |     b = injury
    44     0     0 |     c = fatalities
```

Non-Weka with OneR:

```
=== Summary ===

Correctly Classified Instances       20954               86.4403 %
Incorrectly Classified Instances      3287               13.5597 %
Kappa statistic                         -0.0007
Mean absolute error                      0.0904
Root mean squared error                  0.3007
Relative absolute error                 57.8279 %
Root relative squared error            107.557  %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.865      1.000    0.927      -0.007   0.500     0.865     crash
                 0.000    0.000    0.000      0.000    0.000      -0.007   0.500     0.133     injury
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     fatalities
Weighted Avg.    0.864    0.865    ?          0.864    ?          ?        0.500     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20954     8     0 |     a = crash
  3235     0     0 |     b = injury
    44     0     0 |     c = fatalities
```

CorrelationAttributeEval with Naive Bayes:

```
=== Summary ===

Correctly Classified Instances       20061               82.7565 %
Incorrectly Classified Instances      4180               17.2435 %
Kappa statistic                          0.1405
Mean absolute error                      0.156
Root mean squared error                  0.2866
Relative absolute error                 99.8195 %
Root relative squared error            102.5406 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.927    0.807    0.880      0.927    0.903      0.145    0.706     0.938     crash
                 0.193    0.071    0.294      0.193    0.233      0.146    0.706     0.254     injury
                 0.023    0.002    0.024      0.023    0.023      0.022    0.745     0.013     fatalities
Weighted Avg.    0.828    0.707    0.800      0.828    0.812      0.145    0.706     0.845

=== Confusion Matrix ===

     a     b     c   <-- classified as
 19437  1492    33 |     a = crash
  2604   623     8 |     b = injury
    41     2     1 |     c = fatalities
```

CorrelationAttributeEval with J48

```
=== Summary ===

Correctly Classified Instances       20961               86.4692 %
Incorrectly Classified Instances      3280               13.5308 %
Kappa statistic                          0.0039
Mean absolute error                      0.1559
Root mean squared error                  0.2794
Relative absolute error                 99.7391 %
Root relative squared error             99.941  %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.997    0.865      1.000    0.927      0.028    0.507     0.867     crash
                 0.003    0.000    0.474      0.003    0.006      0.028    0.507     0.138     injury
                 0.000    0.000    ?          0.000    ?          ?        0.474     0.002     fatalities
Weighted Avg.    0.865    0.862    ?          0.865    ?          ?        0.507     0.768

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20952    10     0 |     a = crash
  3226     9     0 |     b = injury
    44     0     0 |     c = fatalities
```

CorrelationAttributeEval with RandomForest

```
=== Summary ===

Correctly Classified Instances       20579               84.8934 %
Incorrectly Classified Instances      3662               15.1066 %
Kappa statistic                          0.0654
Mean absolute error                      0.1423
Root mean squared error                  0.2835
Relative absolute error                 91.0186 %
Root relative squared error            101.4163 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.970    0.925    0.870      0.970    0.917      0.083    0.674     0.922     crash
                 0.075    0.030    0.281      0.075    0.119      0.083    0.674     0.222     injury
                 0.000    0.000    0.000      0.000    0.000      -0.001   0.585     0.012     fatalities
Weighted Avg.    0.849    0.804    0.790      0.849    0.809      0.083    0.674     0.827

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20336   620     6 |     a = crash
  2991   243     1 |     b = injury
    41     3     0 |     c = fatalities
```

CorrelationAttributeEval with DecisionTable

```
=== Summary ===

Correctly Classified Instances       20960               86.4651 %
Incorrectly Classified Instances      3281               13.5349 %
Kappa statistic                          0.0326
Mean absolute error                      0.1481
Root mean squared error                  0.2703
Relative absolute error                 94.7638 %
Root relative squared error             96.6925 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.977    0.867      0.996    0.927      0.084    0.707     0.938     crash
                 0.023    0.004    0.494      0.023    0.045      0.085    0.710     0.261     injury
                 0.000    0.000    ?          0.000    ?          ?        0.679     0.006     fatalities
Weighted Avg.    0.865    0.845    ?          0.865    ?          ?        0.708     0.846

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20884    78     0 |     a = crash
  3159    76     0 |     b = injury
    44     0     0 |     c = fatalities
```

CorrelationAttributeEval with OneR

```
=== Summary ===

Correctly Classified Instances       20951               86.428  %
Incorrectly Classified Instances      3290               13.572  %
Kappa statistic                          0.0009
Mean absolute error                      0.0905
Root mean squared error                  0.3008
Relative absolute error                 57.8807 %
Root relative squared error            107.6061 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.999    0.865      0.999    0.927      0.006    0.500     0.865     crash
                 0.001    0.001    0.211      0.001    0.002      0.006    0.500     0.134     injury
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     fatalities
Weighted Avg.    0.864    0.864    ?          0.864    ?          ?        0.500     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20947    15     0 |     a = crash
  3231     4     0 |     b = injury
    44     0     0 |     c = fatalities
```

InfoGainAttributeEval with Naive Bayes:

```
=== Summary ===

Correctly Classified Instances       20618               85.0542 %
Incorrectly Classified Instances      3623               14.9458 %
Kappa statistic                          0.0914
Mean absolute error                      0.1459
Root mean squared error                  0.2816
Relative absolute error                 93.3577 %
Root relative squared error            100.7449 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969    0.905    0.873      0.969    0.918      0.112    0.701     0.934     crash
                 0.095    0.030    0.328      0.095    0.148      0.115    0.705     0.251     injury
                 0.000    0.001    0.000      0.000    0.000      -0.001   0.691     0.004     fatalities
Weighted Avg.    0.851    0.786    0.798      0.851    0.814      0.112    0.702     0.841

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20310   628    24 |     a = crash
  2925   308     2 |     b = injury
    42     2     0 |     c = fatalities
```

InfoGainAttributeEval with J48:

```
=== Summary ===

Correctly Classified Instances       20962               86.4733 %
Incorrectly Classified Instances      3279               13.5267 %
Kappa statistic                          0
Mean absolute error                      0.1563
Root mean squared error                  0.2795
Relative absolute error                 99.9747 %
Root relative squared error            100      %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.865      1.000    0.927      ?        0.500     0.865     crash
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.133     injury
                 0.000    0.000    ?          0.000    ?          ?        0.473     0.002     fatalities
Weighted Avg.    0.865    0.865    ?          0.865    ?          ?        0.500     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20962     0     0 |     a = crash
  3235     0     0 |     b = injury
    44     0     0 |     c = fatalities
```

InfoGainAttributeEval with RandomForest:

```
=== Summary ===

Correctly Classified Instances       20784               85.739  %
Incorrectly Classified Instances      3457               14.261  %
Kappa statistic                          0.0649
Mean absolute error                      0.1423
Root mean squared error                  0.2765
Relative absolute error                 91.0144 %
Root relative squared error             98.904  %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.940    0.870      0.982    0.923      0.096    0.693     0.928     crash
                 0.060    0.018    0.343      0.060    0.103      0.096    0.695     0.242     injury
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.582     0.003     fatalities
Weighted Avg.    0.857    0.815    0.798      0.857    0.811      0.096    0.693     0.834

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20589   372     1 |     a = crash
  3038   195     2 |     b = injury
    43     1     0 |     c = fatalities
```

InfoGainAttributeEval with DecisionTable:

```
=== Summary ===

Correctly Classified Instances       20957               86.4527 %
Incorrectly Classified Instances      3284               13.5473 %
Kappa statistic                          0.0336
Mean absolute error                      0.1488
Root mean squared error                  0.2713
Relative absolute error                 95.2127 %
Root relative squared error             97.0375 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.976    0.867      0.996    0.927      0.084    0.696     0.933     crash
                 0.024    0.004    0.485      0.024    0.046      0.085    0.700     0.257     injury
                 0.000    0.000    ?          0.000    ?          ?        0.699     0.004     fatalities
Weighted Avg.    0.865    0.844    ?          0.865    ?          ?        0.696     0.841

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20878    84     0 |     a = crash
  3156    79     0 |     b = injury
    44     0     0 |     c = fatalities
```

InfoGainAttributeEval with OneR:

```
=== Summary ===

Correctly Classified Instances       20953               86.4362 %
Incorrectly Classified Instances      3288               13.5638 %
Kappa statistic                          0.0028
Mean absolute error                      0.0904
Root mean squared error                  0.3007
Relative absolute error                 57.8455 %
Root relative squared error            107.5734 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.998    0.865      0.999    0.927      0.017    0.501     0.865     crash
                 0.002    0.001    0.320      0.002    0.005      0.018    0.501     0.134     injury
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     fatalities
Weighted Avg.    0.864    0.863    ?          0.864    ?          ?        0.501     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20945    17     0 |     a = crash
  3227     8     0 |     b = injury
    44     0     0 |     c = fatalities
```

ReliefF with Naive Bayes:

```
=== Summary ===

Correctly Classified Instances       20469               84.4396 %
Incorrectly Classified Instances      3772               15.5604 %
Kappa statistic                          0.1182
Mean absolute error                      0.1467
Root mean squared error                  0.2821
Relative absolute error                 93.8569 %
Root relative squared error            100.9044 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.956    0.867    0.876      0.956    0.914      0.133    0.704     0.935     crash
                 0.133    0.043    0.322      0.133    0.189      0.134    0.707     0.253     injury
                 0.000    0.001    0.000      0.000    0.000      -0.001   0.698     0.004     fatalities
Weighted Avg.    0.844    0.755    0.800      0.844    0.816      0.132    0.704     0.843

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20038   904    20 |     a = crash
  2800   431     4 |     b = injury
    42     2     0 |     c = fatalities
```

ReliefF with J48:

```
=== Summary ===

Correctly Classified Instances       20959               86.461  %
Incorrectly Classified Instances      3282               13.539  %
Kappa statistic                          0.0271
Mean absolute error                      0.1536
Root mean squared error                  0.278
Relative absolute error                 98.237  %
Root relative squared error             99.4597 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.997    0.981    0.867      0.997    0.927      0.076    0.537     0.873     crash
                 0.019    0.003    0.488      0.019    0.037      0.076    0.537     0.170     injury
                 0.000    0.000    ?          0.000    ?          ?        0.482     0.002     fatalities
Weighted Avg.    0.865    0.849    ?          0.865    ?          ?        0.537     0.777

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20896    66     0 |     a = crash
  3172    63     0 |     b = injury
    44     0     0 |     c = fatalities
```

ReliefF with RandomForest:

```
=== Summary ===

Correctly Classified Instances       20440               84.32   %
Incorrectly Classified Instances      3801               15.68   %
Kappa statistic                          0.0679
Mean absolute error                      0.1441
Root mean squared error                  0.2882
Relative absolute error                 92.1724 %
Root relative squared error            103.1132 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.962    0.913    0.871      0.962    0.914      0.080    0.663     0.922     crash
                 0.088    0.038    0.262      0.088    0.132      0.082    0.666     0.209     injury
                 0.000    0.000    0.000      0.000    0.000      -0.001   0.608     0.003     fatalities
Weighted Avg.    0.843    0.795    0.788      0.843    0.808      0.080    0.663     0.825

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20155   803     4 |     a = crash
  2950   285     0 |     b = injury
    44     0     0 |     c = fatalities
```

ReliefF with DecisionTable:

```
=== Summary ===

Correctly Classified Instances       20952               86.4321 %
Incorrectly Classified Instances      3289               13.5679 %
Kappa statistic                          0.0158
Mean absolute error                      0.151
Root mean squared error                  0.2729
Relative absolute error                 96.6034 %
Root relative squared error             97.62   %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.988    0.866      0.998    0.927      0.053    0.679     0.930     crash
                 0.012    0.002    0.442      0.012    0.023      0.054    0.683     0.239     injury
                 0.000    0.000    ?          0.000    ?          ?        0.660     0.003     fatalities
Weighted Avg.    0.864    0.855    ?          0.864    ?          ?        0.680     0.836

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20914    48     0 |     a = crash
  3197    38     0 |     b = injury
    44     0     0 |     c = fatalities
```

ReliefF with OneR:

```
=== Summary ===

Correctly Classified Instances       20953               86.4362 %
Incorrectly Classified Instances      3288               13.5638 %
Kappa statistic                          0.0028
Mean absolute error                      0.0904
Root mean squared error                  0.3007
Relative absolute error                 57.8455 %
Root relative squared error            107.5734 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.998    0.865      0.999    0.927      0.017    0.501     0.865     crash
                 0.002    0.001    0.320      0.002    0.005      0.018    0.501     0.134     injury
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     fatalities
Weighted Avg.    0.864    0.863    ?          0.864    ?          ?        0.501     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20945    17     0 |     a = crash
  3227     8     0 |     b = injury
    44     0     0 |     c = fatalities
```

CfsSubsetEval with Naive Bayes:

```
=== Summary ===

Correctly Classified Instances       20615               85.0419 %
Incorrectly Classified Instances      3626               14.9581 %
Kappa statistic                          0.0929
Mean absolute error                      0.1461
Root mean squared error                  0.2817
Relative absolute error                 93.4582 %
Root relative squared error            100.7558 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969    0.903    0.873      0.969    0.918      0.114    0.705     0.937     crash
                 0.096    0.030    0.329      0.096    0.148      0.116    0.705     0.251     injury
                 0.023    0.001    0.029      0.023    0.025      0.024    0.721     0.011     fatalities
Weighted Avg.    0.850    0.785    0.799      0.850    0.814      0.114    0.705     0.844

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20304   629    29 |     a = crash
  2920   310     5 |     b = injury
    41     2     1 |     c = fatalities
```

CfsSubsetEval with J48:

```
=== Summary ===

Correctly Classified Instances       20962               86.4733 %
Incorrectly Classified Instances      3279               13.5267 %
Kappa statistic                          0
Mean absolute error                      0.1563
Root mean squared error                  0.2795
Relative absolute error                 99.9747 %
Root relative squared error            100      %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.865      1.000    0.927      ?        0.500     0.865     crash
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.133     injury
                 0.000    0.000    ?          0.000    ?          ?        0.473     0.002     fatalities
Weighted Avg.    0.865    0.865    ?          0.865    ?          ?        0.500     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20962     0     0 |     a = crash
  3235     0     0 |     b = injury
    44     0     0 |     c = fatalities
```

CfsSubsetEval with RandomForest:

```
=== Summary ===

Correctly Classified Instances       20790               85.7638 %
Incorrectly Classified Instances      3451               14.2362 %
Kappa statistic                          0.0752
Mean absolute error                      0.1413
Root mean squared error                  0.2759
Relative absolute error                 90.3841 %
Root relative squared error             98.7074 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.931    0.871      0.981    0.923      0.108    0.699     0.928     crash
                 0.068    0.019    0.360      0.068    0.115      0.107    0.699     0.248     injury
                 0.000    0.000    0.000      0.000    0.000      -0.001   0.602     0.011     fatalities
Weighted Avg.    0.858    0.808    0.801      0.858    0.813      0.108    0.699     0.836

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20569   391     2 |     a = crash
  3012   221     2 |     b = injury
    42     2     0 |     c = fatalities
```

CfsSubsetEval with DecisionTable:

```
=== Summary ===

Correctly Classified Instances       20957               86.4527 %
Incorrectly Classified Instances      3284               13.5473 %
Kappa statistic                          0.0314
Mean absolute error                      0.148
Root mean squared error                  0.2702
Relative absolute error                 94.695  %
Root relative squared error             96.6659 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.996    0.977    0.867      0.996    0.927      0.082    0.710     0.939     crash
                 0.023    0.004    0.483      0.023    0.043      0.081    0.713     0.263     injury
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.674     0.006     fatalities
Weighted Avg.    0.865    0.846    0.814      0.865    0.807      0.081    0.710     0.847

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20884    78     0 |     a = crash
  3161    73     1 |     b = injury
    44     0     0 |     c = fatalities
```

CfsSubsetEval with OneR:

```
=== Summary ===

Correctly Classified Instances       20951               86.428  %
Incorrectly Classified Instances      3290               13.572  %
Kappa statistic                          0.0009
Mean absolute error                      0.0905
Root mean squared error                  0.3008
Relative absolute error                 57.8807 %
Root relative squared error            107.6061 %
Total Number of Instances            24241

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.999    0.999    0.865      0.999    0.927      0.006    0.500     0.865     crash
                 0.001    0.001    0.211      0.001    0.002      0.006    0.500     0.134     injury
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.002     fatalities
Weighted Avg.    0.864    0.864    ?          0.864    ?          ?        0.500     0.766

=== Confusion Matrix ===

     a     b     c   <-- classified as
 20947    15     0 |     a = crash
  3231     4     0 |     b = injury
    44     0     0 |     c = fatalities
```

Part 5: Discussion and Conclusion

Section 5.1: Looking at Results

Our best combination of attribute analysis and a classifier model was CfsSubsetEval with Naive Bayes, achieving accuracy of 85.0519%, TP rate of 0.850, FP Rate of 0.785, precision of 0.799, recall of 0.850, f-measure of 0.814, and MCC of 0.114. We chose this as our best model because it had the highest f-measure and MCC, which considers the parts of the confusion matrix and precision and recall. However, there were other combinations of attribute selection algorithms and a classifier that achieved better performance in the metric of accuracy. The tradeoff between accuracy and precision and recall in this scenario is largely due to how imbalanced the class labels are in the dataset. This leads to the model preferring to classify instances as “crash” since there are few “injury” and “fatalities” labels. High accuracy will come from correctly classifying the crashes as crashes, but this often results in misclassifying the instances with “injury” and “fatalities” class labels. We thought rather than being extremely biased towards the majority class and always predicting “crash”, it would be better to measure the model’s usefulness in predicting the other classes by considering non-accuracy metrics.

Section 5.2: Future Work

In the future, we can work on addressing the challenges posed by the imbalanced class distribution within the dataset. One promising avenue is to explore advanced sampling techniques, such as synthetic minority over-sampling or under-sampling methods, to create a more balanced dataset for training. We attempted to use WEKA’s method of balancing classes by adjusting the weights, but this resulted in the majority of the feature selection algorithms being unusable. We can try utilizing other methods online that directly artificially add or remove instances instead of changing WEKA weights.

Section 5.3: Team Member Roles

Finding the Data & Dataset Overview: Gabriel
Preprocessing: Gabriel
Attribute Analysis: Andrew
Classifier Models: Andrew
Discussion and Conclusion: Gabriel